

Research is no substitute for political action on climate

Tax incentives and other measures to encourage energy efficiency are needed to curtail the growth of US carbon emissions. Calls for more research, though important, should not be used to obscure this fact.

The idea of a government-sponsored or international initiative to develop non-carbon energy sources, proposed last week by Hoffert *et al.* (*Nature* 395, 881–884; 1998), should sound alarms in one respect, at least: results of previous energy technology drives have not always been encouraging. Governments invested billions of dollars in the research, development and demonstration of nuclear power in the 1950s and 1960s, but ultimately the technology failed as an economically competitive energy source when its full life-cycle costs were taken into account. In the United States twenty years ago, significant funds were directed at renewable energy by the Carter administration. Much effort was wasted on 'demonstration projects' that featured little genuine innovation. With these histories in mind, there is all the more onus on its proponents to show that the proposed energy R&D would work out differently.

A year ago, a panel of the President's Committee of Advisors on Science and Technology, chaired by John Holdren of Harvard University, made the case for an expanded energy R&D programme. Their report prudently stressed the importance of a broad research portfolio. Holdren would not argue that his recommendations would greatly influence US energy use in the near term. Yet that is the claim that the Clinton administration is now making for the far smaller research package that it extracted from recent budget negotiations.

In fact, technology support, while important, is only one of the things that the United States should be doing to curtail its emissions. Encouraging energy efficiency is not rocket science. It merely requires a little political courage, which has not been forthcoming thus far from President Clinton. He has proposed, for example, a tax incentive to encourage people who buy big cars to buy big cars with better fuel economy instead. An incentive to encourage the purchase of small cars was rejected as too rude to Detroit,

which can't produce small cars efficiently.

Clinton may be cautious, but his opponents in the Congress make him look like a reckless adventurer. For example, conservatives in the House of Representatives have been obstructing the introduction of federal standards for the efficiency of domestic appliances. They base their objection on the view that a federal government notice on a washing machine, announcing its average wattage, is one step too far down the slippery slope to state intervention in the lives of ordinary Americans. Perhaps its real basis is more to do with makers of inefficient washing machines, who find political contributions to be less financially exacting than modern industrial design.

Where the political will exists, significant emissions cuts can be made today, without recourse to exotic technology. In this regard, it is encouraging to note that Senator Connie Mack (Republican, Florida) is supporting a bipartisan measure that would prepare the ground for future tax credits for corporations that take action to reduce emissions now (see page 7). Mack is a staunchly conservative senator, but most of the people in his state live barely a few feet above sea level, and may be reluctant to participate in a lengthy experiment to establish whether political paralysis is an adequate response to the mounting scientific evidence for man-made climate change.

There is a role for government in supporting the scientific research needed to underpin a healthy, climate-neutral energy supply industry. It will then be largely up to that industry to implement the technologies that will cut carbon emissions, and it will only be tempted to do so when the right price and tax incentives exist. For now, when governments talk of research programmes to counteract climate change, there is a danger that they are placing a fig-leaf over their own failure to put such incentives in place. □

Reform in name only

Using threats to force the formation of semi-autonomous 'agencies' in Japan is a flawed strategy.

In bulldozer fashion, Japan's Prime Minister Keizo Obuchi is trying to force through plans to convert large numbers of national research institutes and possibly some universities into semi-autonomous corporations or 'agencies' by threatening to cut the budgets of those that do not comply by 30% (see page 7). If Obuchi's zeal were driven primarily by a proper understanding of the need to make Japan's universities and national institutes more accountable, more independent and more performance driven, his move should be applauded. Sadly, that does not seem to be the case.

Rather, his prime motivation seems to be to go one up in the public eye over his more charismatic predecessor Ryutaro Hashimoto, who initiated the administrative reform, by converting as many as 100 government institutions into 'agencies', so that the new administration can claim to have trimmed the government significantly.

So determined is Obuchi to achieve this goal that national institutions are reportedly being offered the carrot that if they convert to 'agencies' they can keep exactly the same budgets and staffing arrangements as they currently have. As one top official from a leading national research institution wryly observes, he would prefer to stay within the government's fiefdom, have the institution's budget cut by 30% and thereby have the grounds to clear out some of the dead wood among the staff that he has wanted to remove for years.

But the cuts seem unlikely to materialize. Preservation of budgets is the *raison d'être* of all bureaucrats in Japan's science-related ministries and agencies. Most government researchers want to maintain the status quo, and even some of Obuchi's colleagues in the ruling Liberal Democratic Party are openly expressing doubts about the reform plan. Much-needed change in Japan's government research system is likely to remain elusive. □



UK's physical scientists are left disappointed by budget choices

[LONDON] Britain's physical scientists expressed both surprise and concern last week at the relatively large proportion of an extra £300 million (US\$500 million) being made available to research councils over the next three years that has been reserved for the life sciences.

In revealing how this extra money, first promised in July, is to be divided between research areas, the Department of Trade and Industry has announced that a key priority is to secure a "major expansion in molecular, biomolecular and biomedical research" (see *Nature* 395, 825; 1998).

In addition to a significant increase in funding for the Medical Research Council, the Engineering and Physical Sciences Research Council (EPSRC) has been told to spend £60 million of its £86 million extra income on research that underpins the work of the biomedical and environmental research councils.

The government is also to invest £35 million in DIAMOND, a planned synchrotron radiation source, in addition to £110 million pledged by the Wellcome Trust. There will also be at least 57 new research fellowships, of which 12 are reserved for women, funded by the Royal Society and the Royal Academy of Engineering.

Research into climate change, ageing, information technology and communications will also be given funding priority, although precise amounts have yet to be spec-



Secure: budget increase guarantees Britain's role in the construction of the Large Hadron Collider.

ified. But funding for particle physics and astronomy research will stay constant in real terms for the next three years.

Senior physicists say they had expected the biomedical and environmental sciences to do well in the distribution. But few expected them to walk away with almost 80 per cent of the extra money, leaving physical scientists and engineers working outside priority areas with £46 million.

Particle physicists and astronomers will receive £20 million of this, an increase in real terms of just over half a per cent. This will be used for, among other things, supporting Britain's contribution to international projects such as the Large Hadron Collider at CERN the European Laboratory for Particle Physics in Geneva. The funds will also be used

for new fellowships for mid-career scientists, and a fund for new and innovative projects.

Other physical scientists whose work has no bearing on the environmental or biological sciences could fare worse. It is not yet known if their extra £26 million represents a real-terms increase on present budgets.

Ken Pounds, head of the department of physics and astronomy at the University of Leicester, and a former chief executive of the Particle Physics and Astronomy Research Council (PPARC), admits that the overall situation has left many physicists unhappy.

"I personally think [keeping spending roughly level in real terms] is good news and reverses a 20-year trend," says Pounds. "But the situation is no better than marking time. It keeps [physics] alive."

Roger Cashmore, chairman of the department of physics at the University of Oxford, describes the distribution between the research councils as "a little peculiar", and thinks it is "somewhat unfortunate" that his own area of particle physics "did not do well at all".

Officials from PPARC and the EPSRC have greeted the extra funds as a long-awaited and much appreciated boost to their incomes.

The previous funding decline had led PPARC to close down the Royal Greenwich Observatory in Cambridge, a telescope building facility, and merge some of its work with the Royal Observatory in Edinburgh, creating a new Astronomy Technology Centre. The closure allows PPARC to invest £11 million in its domestic programme.

PPARC and EPSRC welcome the chance to plan their expenditure over three years instead of just one. PPARC officials are also pleased by the decision of the government's Office of Science and Technology to hold an additional £30 million to guarantee Britain's contribution to existing international projects, such as the European Space Agency, in the event of currency fluctuations.

But privately their relief is said to be tempered with some concern. No one wants to complain, as that might seem ungrateful. But the feeling in the community, says one official, is: "When you give a starving man a slice of bread, he's really happy; but when he sees his chums eating cake, it hurts," he says. "That's how some of us feel right now."

Some officials are concerned that the imbalance in the distribution of the research funds may also be reflected in the distribution of an additional £600 million fund for research infrastructure, half of which is being provided by the Wellcome Trust (see *Nature*

Dutch universities keep research control



Hermans: Easing the universities' burden.

[AMSTERDAM] The planned transfer of a significant proportion of the research budget of Dutch universities to the national basic research organization, NWO, has been abandoned. The decision was announced by the new research minister Loek Hermans (left), who took office in August

following the re-election of a Liberal-Labour coalition government.

The transfer of Dfl 500 million (US\$267 million), about 20 per cent of the universities' research budget, had been planned last year by the previous government (see *Nature* 390, 9; 1997), which was also a coalition of Liberal and Labour

parties. Unsurprisingly, the scheme had been championed by the NWO but opposed by universities, where most Dutch basic research is carried out.

Hermans, a professional politician with no research experience, told universities two weeks ago that he was abandoning the scheme because it would place too great a burden on them. They have recently been told that their general budget will be reduced in stages by Dfl 285 million over the next few years (see *Nature* 394, 405; 1998).

The NWO has not yet received formal notification of Hermans' decision. But Reinder van Duinen, NWO's president, describes the decision as "a shame" because he believes that the proportion of basic research money allocated through competitive grants is too low. To run a healthy dual funding system in the Netherlands, the amount of competitive grant money should be doubled, he says.

Alison Abbott

395, 422; 1998).

The trust's status as a biomedical research charity means that at least £300 million will be spent on biomedical research. But there are concerns that the overall proportion may be much greater. Indeed, questions have been asked as to whether the trust's involvement will skew the allocation of the overall fund towards the life sciences.

Asked to address this point by members of the House of Commons' science and technology committee last week, Mike Dexter, director of the trust, said that there were no strings attached to the overall funds. Indeed, Sainsbury, the science minister, said the trust's contribution to research infrastructure funding meant that, in total, the government was able to devote more funds to the physical sciences than it might otherwise have done.

Although there is disagreement over the proportion of research funds reserved for the life sciences, few doubted that life sciences would receive more than the physical sciences. Not only does it reflect a similar development in the United States, but it also ties in with the government's stated priorities to fund research that enhances its social and economic goals.



Cashmore: division The government
was 'a little peculiar'. firmly takes the view
that science is the

bedrock of economic success. One of its stated goals, announced with the three-year allocations, is that the public-sector science base should contribute to a 50 per cent increase annually in the number of companies set up.

The government is aware that Britain is home to some of the world's leading pharmaceutical and biotechnology companies. Partly because of this, much of the investment in the life sciences will focus on exploiting the data to emerge from the Human Genome Project.

According to some, the economic potential of research, particularly in the life sciences, was a key factor on which Britain's finance ministry, the Treasury, agreed to release the extra funds in the first place.

The requirement that the EPSRC uses much of its new money to support the work in the life and environmental sciences is welcomed by one senior research council official as an opportunity for physical scientists and engineers to use their skills in other fields.

But he believes the government may be putting too much faith in the idea that high-quality research is essential for commercial success. Japan's prosperous automobile and electronics companies succeeded not by having the world's best research, he says, but primarily by having higher productivity than Britain.

Ehsan Masood

Synchrotron report backs steps to efficient usage

[MUNICH] National synchrotron facilities in Europe need to become more efficient, both technically and administratively, if they are to keep up with the increasing demand from life scientists, according to a report published by the European Science Foundation (ESF) two weeks ago.

Indeed, the report warns that unless this happens, an inadequate availability of synchrotron radiation facilities could hold back efforts to understand the function of novel proteins identified by genome sequencing programmes.

Synchrotron radiation has become an essential tool in protein crystallography. Structural information can be inferred from the patterns of X-ray scattering and absorption that result from impact with the molecules in a crystal, and synchrotron sources of X-rays can be tuned to different wavelengths, allowing a wide range of proteins to be analysed.

The third-generation European Synchrotron Radiation Facility (ESRF) in Grenoble is particularly valued for the intensity of its X-rays, which allows bigger molecules to be analysed with less damage to the crystal. A further eight national facilities around Europe also provide dedicated beam lines for life scientists, with four more under construction and two in the proposal stage.

The ESF report endorses the need for the two proposed facilities, one near Barcelona in Spain and the other, named Soleil, in France. But it admits that any further proposals are unlikely to win financial support from national governments in the current economic climate.

Already the heavy demand means that scientists must often wait many months for beam time. This can cause problems for protein X-ray crystallography because of the instability of most protein crystals.

But the report says that such delays are often avoidable. Indeed, rather than recommending the construction of further synchrotron sources, it says existing ones should be made to function more efficiently, and makes five specific recommendations for increasing efficiency.

First, all beam lines should be equipped with CCD detectors — highly efficient X-ray detectors — to speed up the read-out of data. The ESRF has installed these on nearly all its beam lines. But despite their relatively low cost and their ability to increase throughput by about tenfold, national facilities have not always received money to make this upgrade.

Second, staffing of beam lines should be improved. Unlike physicists using synchrotron radiation, life scientists are often inex-

perienced and need more help. But good beam-line operators are hard to find, and the report says that such positions should be made more attractive by allowing operators time to pursue their own research.

Third, and most importantly, the report says application procedures for beam-line time should be adjusted to meet the needs of the life sciences community. Many European scientists using synchrotron facilities in other countries are funded by the European Commission in Brussels. But the long procedure for peer reviewing commission research project applications introduces impractical delays for protein crystallographers.

The report therefore recommends a 'twin-track' system of time allocation: a block booking for highly qualified research groups, and a fast-track system, giving access to beam lines within a few days, for individual users with crystals that are difficult to obtain or store.

Block booking would mean a change in the peer-review system for individual projects, and scientists acknowledge that this is a sensitive issue. "There is obviously a need to be seen to be spending public money in an accountable way", says Keith Wilson, professor of chemistry at the University of York in the United Kingdom, and vice-chairman of the report's expert review panel.

Other protein crystallographers admit that synchrotron beam time is so precious, and the competition to publish so fierce, that researchers often run several projects in parallel, using beam time allocated for a particular peer-reviewed project for whichever of the projects has come to fruition at the appointed time. They also routinely send essentially the same applications to different facilities in Europe to increase the chances of securing time.

The ESF report's fourth suggestion is that synchrotron facilities consider providing services for collecting data and assessing the quality of crystals, to save expense and time bringing scientists to the site.

Finally, it suggests creating a committee of European synchrotron users and providers in the life sciences, administered by the ESF, and equivalent to the US organization Biosync. Such an umbrella organization could help coordinate and monitor the use of European facilities to identify problems and bottlenecks hampering efficient use.

Simon Phillips, an X-ray crystallographer from the UK's University of Leeds, agrees that the recommendations "make a lot of sense. Interesting crystals can be generated very suddenly, and it is frustrating to have to wait up to a year for beam time."

Alison Abbott

Gore wins approval for his Triana satellite

[WASHINGTON] The US space agency NASA and Congress have given the go-ahead to a \$75 million satellite that will regularly return updated pictures of Earth from space, beginning in early 2001.

Dubbed 'Triana' after the Spanish sailor who was Columbus's lookout, the satellite was conceived by Vice-President Al Gore as an "inspirational" project (see *Nature* 394, 213; 1998). Money for Triana will come from the agency's existing \$1.4 billion budget for Earth science.

A mission concept by Francisco Valero of the Scripps Institution of Oceanography in San Diego has beaten eight other proposals submitted to NASA earlier this year. Triana will be stationed at the L1 'neutral gravity' Lagrange point, 1.5 million kilometres from Earth in the direction of the Sun. It will have a 16-channel, multispectral imaging camera designed by Lockheed-Martin, which has flown previously on other spacecraft. Valero's group also plans to fly a radiometer for measuring the total energy reflected from the Earth, and a small solar wind monitor.

Measuring Earth's reflected energy currently requires piecing together thousands of satellite observations, says Valero. Triana's whole-Earth view will allow scientists to "look at the planet in a new way". Its unique viewing geometry will also show "hot spots" of reflected light, yielding information about forest canopy structure and vegetation health.

Owen Toon of the University of Colorado, who heads one of the interdisciplinary science teams for NASA's Earth Observing System (EOS) of satellites scheduled to



Mosaic: this view of Earth was assembled from 70 images. Triana will capture the scene in one frame.

start launching next year, believes Triana will be an "interesting platform" for observing Earth from a new location. Piecing together the global radiation budget is not a negligible problem, he says.

NASA had no plans to fly an all-Earth radiometer before Gore came up with the Triana concept. Bruce Wielicki of NASA's Langley Research Center in Virginia, the principal science investigator for the EOS CERES instrument that will measure energy reflected from clouds, agrees that the new vantage point could have scientific value. But he believes, like many scientists, that Triana will be "more of an emotional and educational thing" than a research programme. "I don't think it's what [scientists] would have picked a priori," he says.

Wielicki, however, believes the Earth science community could use a high-profile mission to engender public interest, just as the Hubble Space Telescope does for space-based astronomy. "The American public, to be honest, doesn't want just pure science," he says.

NASA, which announced its choice of Valero's design last week, is not saying what other parts of the Earth science programme will be squeezed to pay for Triana, although it has submitted to Congress a plan for "rephasing" some projects. Building the spacecraft and instruments will cost \$35 million this fiscal year, with the balance spent in 2000.

Although many researchers are uneasy with the way Triana came about — Gore effectively handed his pet project to NASA as an assignment — it is unlikely to be a major burden on the Earth science community. "It's a small enough project that everybody knows it's not going to break the bank," says Wielicki.

The House of Representatives prevented NASA from spending money on Triana until Congress was convinced it had support in the scientific and/or commercial sectors. But that restriction was removed in the agency's funding bill passed last month. Congressional sources say Gore took his case in person to Jerry Lewis (Republican, California), the House appropriations subcommittee chairman, before the bill was finalized.

Valero's concept leaves the non-scientific aspects of the mission to NASA. The agency says it plans to solicit educational proposals next year and commercial partnerships "over the coming months". **Tony Reichhardt**

Berkeley teams up with Novartis in \$50m plant genomics deal

[WASHINGTON] The University of California at Berkeley (UCB) is poised to strike a unique, \$50 million agreement with the Swiss drug company Novartis. The result will be that a research institute linked to the company will have first rights to a proportion of the genomics research conducted at Berkeley's Department of Plant and Microbial Biology.

The deal is expected to bring the department up to \$25 million for laboratory refurbishment and a further \$25 million in research support over five years, and could be signed by both parties as early as next week. It will differ from normal collaborative agreements between universities and industry because it will involve the entire department, which has 25 professors.

Berkeley will have access to gene-sequencing technology and database information on plant genomics that it would otherwise be impossible for university researchers to obtain. In return, Novartis

Agricultural Discovery Institute, Inc. (NADII), a major new plant research centre being built at La Jolla, California, will get first rights to negotiate for between a quarter and a third of all the discoveries made at the department.

That fraction corresponds approximately to the proportion of the department's total research budget that will be provided by NADII. The institute is entirely funded by Novartis through a charitable foundation, although the two entities are legally independent.

According to a spokesman for UCB, the university's plant and microbial department has been seeking a corporate partner for some time, and spoke with four contenders before negotiating the deal with Novartis.

The deal has already drawn criticism from some students and staff at UCB. They say that it was drawn up in secret and could threaten academic freedom at the university. But Bob Buchanan, the chairman of the

department, argues that "it doesn't impinge on academic freedom at all".

Buchanan adds that the arrangement will bring the department the resources it needs to support graduate students and postdoctoral fellows in the way that medical schools do. "The driving force for this is not the research money, but the training aspect. It will give students opportunities that we couldn't otherwise give them."

Steven Briggs, president of NADII, says it is also seeking "one more academic alliance of similar scope" to that about to be forged with UCB.

NADII was established earlier this year, and expects to move into a new facility with up to 200 scientific staff in two years' time. The research centre is one of a spate of new ventures into plant genomics by corporations. They expect that rapid progress in the discipline will revolutionize not just agriculture but also food production and nutrition. **Colin Macilwain**

Asian states link to popularize science...

[ISLAMABAD] Seven south Asian states, including some of the world's poorest countries, have agreed to pool their ideas and expertise to raise the public awareness of science in their region.

Representatives from countries including India, Pakistan and Sri Lanka agreed at the end of a five-day workshop on science popularization last month to set up a regular forum to exchange experiences of incorporating science into programmes on public health, literacy and rural development.

They pledged to hold an annual regional science festival, set up a student exchange programme and investigate setting up a peer-reviewed journal of science and development communication in south Asian countries.

The meeting also recommended that state-controlled media networks be required to allocate ten per cent of air-time to science-related programmes, and that one per cent of this should be during peak audience time.

The meeting, held in Islamabad, was organized by the Pakistan Science Foundation and the South Asian Association for Regional Cooperation (SAARC), an inter-governmental body made up of India, Pakistan, Bangladesh, Sri Lanka, Nepal, Maldives and Bhutan.

SAARC exists partly to improve trade,

cultural and scientific links between member states. Some SAARC countries, however, although neighbours, have tense relations, and delegates expressed concern that these differences may hinder practical cooperation.

For example, India and Pakistan cannot exchange television and radio programmes on science because broadcasts from each country are banned in the other.

Many Pakistani delegates said they considered India's approach to the public understanding of science more relevant to Pakistan's needs. But they said they could never say this in public as it would be too politically sensitive. SAARC governments do not want India to be seen as the region's superpower.

In general, India and Pakistan approach the popularization of science with different aims. Pakistan has adopted what is known as the 'deficit' model, which focuses on helping people to acquire more knowledge about modern scientific developments.

Pakistan's science planners have modelled their country's science promotion activities on those of developed countries. They are based around museums, planetariums and mobile exhibitions, and assume a single target audience who can read and write.

India, on the other hand, has adopted a more indigenous approach. For example,

planetariums are rarely used in promoting an understanding of astronomy. Children are encouraged instead to build their own telescopes, even down to grinding the mirrors.

"A planetarium is useful in European countries where the sky tends to be cloudy. But in India we have clear night skies for most of the year," says B. K. Tyagi of India's National Council for Science and Technology Communication.

India has also tailored its activities according to its audience. Some of its science popularization activities are aimed only at literate people, and others at those unable to read and write. A third category — which includes campaigns against smoking, drugs and alcohol — targets a general audience.

Science promotion among illiterate rural populations is largely information based. For example, people are taught how to weigh and measure, how to test water quality, and how to become blood donors.

Those who can read and write are encouraged to acquire research skills in addition to facts. They are also given opportunities to use these skills to tackle important local concerns, such as reducing air pollution and improving the quality of the water supply (see box).

Delegates at the workshop recalled many experiences that were common across the SAARC region. These include the problem of communicating science without challenging concepts of religious faith, which remains strong in all SAARC countries.

Science communicators in some SAARC countries are hesitant about raising issues they think will bring them into conflict with religious beliefs. This is particularly true of Pakistan, a Muslim country, where human evolution is rarely discussed in public.

One Pakistan Television executive at the workshop said that, when covering the issue of cloning, he had to take care that the network could not be seen to be promoting the idea.

Delegates were generally critical of the media in their countries. The exception was All India Radio, which regularly broadcasts science programmes. The print media was criticized for ignoring local science and instead publishing syndicated articles from international news agencies.

Delegates spoke of the marked interest shown by ordinary citizens in defence science and technology. This could be a consequence of the region's troubled politics. For example, the media in India and Pakistan devote much space to nuclear issues.

Aleem Ahmed, the editor of the Urdu-language science monthly *Global Science*, which is based in Karachi, says that editions of the magazine that focus on defence or nuclear matters outsell those on other topics.

Ehsan Masood

...and India's young get a taste for research

[ISLAMABAD] Next January, while the senior figures of Indian science gather for the National Science Congress in Madras, the city will be hosting a parallel science congress for children.

For the past five years, India's National Council for Science and Technology Communication (NCSTC) has been experimenting with an unusual method of science communication. Children from schools in almost every district of India are being encouraged to undertake research projects. The best are shown at the annual children's congress.

The idea, says B. K. Tyagi of the NCSTC, is to give children a feel for how science works, to encourage them to consult experts, "and to show them that they can use the tools of research to change their own lives".

This year, some 11,000 schools from 450 districts are



Tapping into schools: children are combating water pollution.

taking part. Projects can be submitted in any of India's 18 official languages. They are assessed according to four criteria: the nature of the problem, its relevance to local communities, the quality of the written report, and the oral presentation.

Tyagi says that, unexpectedly, schools in

rural areas come up with some of the more innovative projects. One 13-year-old identified and eliminated the source of contamination in his village stream.

Elsewhere, children monitored air-pollution levels and persuaded the local authorities to take remedial action.

E.M.

US and Europe set to clash over emissions

[WASHINGTON] Conflict between Europe and the United States over the use of emissions trading to meet targets for cuts in carbon emissions is set to overshadow the opening this week of the fourth Conference of the Parties to the International Framework Convention on Climate Change in Buenos Aires.

The European Union will seek to place restrictions on the extent to which nations can rely on trading to achieve the cuts they agreed to at the last conference in Kyoto, Japan. It has been encouraged in this by environmental groups that believe the United States wants to use emissions trade with Russia and developing countries to avoid any significant cuts to its own emissions.

The meeting of 175 countries in Argentina, which began on 2 November, will try to agree on a timetable for resolving how emissions trading should work, as well as other critical details of the protocol agreed at Kyoto last December. These include the treatment of carbon 'sinks', the proposed Clean Development Mechanism to help control emissions in developing countries, and compliance mechanisms to enforce the protocol.

But none of these issues will be settled at Buenos Aires, where protagonists are expected instead to argue about the timetable for their settlement. Some nations will push for a strict timetable, with most of the points of contention being resolved by the sixth conference of the parties, due to take place in 2000 or 2001.

The developing countries, led by China and India and represented by the so-called Group of 77 (G77), are expected to use their voting power to knock the issue of setting fixed emission targets off the agenda at the meeting. The US delegation will try to get it reinstated, but there is no expectation of any significant movement towards such targets for developing countries.

Instead, the United States and its allies are looking to Carlos Menem, the Argentinian president, to provide a semblance of progress on the issue by promising further voluntary cuts in emissions by Argentina when he addresses the conference.

Late last week, the Clinton administration was still to decide whether Al Gore, the vice-president, would go to the meeting. Environmental groups were hoping that Gore would not only attend but would also sign the Kyoto Protocol. This would be a largely symbolic gesture, however, since there is no prospect of Clinton submitting the measure for the necessary ratification by the US Senate.

But in the run up to Buenos Aires, there has been some indication of reduced hostility in the Senate to cuts in US carbon emissions. Last month, Senator Connie Mack (Republican, Florida) co-sponsored a bill to

give corporations credits for emissions cuts they make now, in advance of any future cuts required to meet the Kyoto targets. This tacit admission by an influential, conservative Republican senator that domestic emissions need to be cut has greatly encouraged supporters of the Kyoto treaty in the United States.

"Senator Mack's action is very significant," says Senator Joe Lieberman (Democrat, Connecticut), a co-sponsor of the measure and a supporter of emissions cuts, who will attend the Buenos Aires meeting. "We're on the right side of history here."

Initiatives to limit emissions "can be frustrated for a short period of time", he adds, predicting that recent announcements by corporations planning to cut their own emissions are "going to wear down opposition to action on climate change".

Last week, a coalition of the companies BP, General Motors and Monsanto, and the World Resources Institute released a report suggesting extensive, voluntary, precautionary action to measure and reduce emissions.

But some environmental groups continue to doubt the intentions of both international corporations and the US delegation at



the Buenos Aires meeting. The United States "is trying every possible means to [avoid] domestic cuts in emissions — and the Europeans are the most powerful folks in their way," says Kalee Krieder, a climate change specialist at Greenpeace.

Dan Lashof of the National Resources Defense Council, another environmental group, argues that "the United States should reassure other governments of our commitment, and outline steps to begin real emissions reductions at home". **Colin Macilwain**

Recalcitrant institutes face cuts in Japan

[TOKYO] Japanese ministries and agencies that do not cooperate with the government's planned administrative reform, including the restructuring of national research institutes and universities, could open such organizations to substantial budget cuts in the next fiscal year as a penalty.

This warning was delivered last week by Hiromu Nonaka, the chief cabinet secretary, in announcing the details of the interim report on administrative reform, which includes recommendations to transform government-run organizations into semi-autonomous bodies.

The reform plan has met strong opposition from ministries and agencies. In particular, national research institutes and universities have shown concern about the effect of such reorganization on the quality of their research (see *Nature* 395, 730; 1998). The interim report also describes their concerns over the difficulty of transforming all the targeted organizations into new corporations with administrative independence.

Frustrated by the ministries' lack of enthusiasm for implementing the plan, Keizo Obuchi, Japan's prime minister, has now told the Management and Coordination Agency to take firm action against institutions resisting the reform

plan by imposing a 30 per cent cut on their budgets for the fiscal year 1999, which begins next April.

"At the moment, there is no feeling of enthusiasm among ministries and agencies to carry out a drastic reorganization of their administrative systems," Nonaka said at a press conference last week. "It is therefore necessary to take strong action in order to push through with the planned reform."

The administrative reform plan is also starting to attract criticism from within the ruling Liberal Democratic Party (LDP). Some LDP politicians, including those involved in administrative reform, are calling for the plan to be postponed so they can prioritize plans to improve the nation's ailing economy.

At a party meeting last week, members of the LDP's administrative reform committee expressed concern about the lack of progress in plans to revive the Japanese economy, arguing that too much effort has been spent on the planned restructuring of government ministries and agencies.

The government hopes to finalize its reorganization plans by next January, but many are predicting considerable delays in the wake of the criticisms of Obuchi's reform efforts. **Asako Saegusa**

'Big science' forum gets a broader role ...

[PARIS] An intergovernmental body, set up six years ago by the Organization for Economic Cooperation and Development (OECD) to act as an international forum for the discussion of issues related to the funding of major science facilities, seems set to be given a new lease of life.

Earlier this year, some member states of the Paris-based OECD were saying that its so-called Megascience Forum might be closed down, partly because it had not been performing what many had seen as its key role. The forum, as its name suggested, was initially supposed to act primarily as a focal point for intergovernmental negotiations on new 'big science' facilities.

But, after a positive review by a four-member external panel, it is now expected

that the forum's work will be continued under a new name — perhaps the Global Science Forum — and with a broader mandate to discuss science policy issues. "We want to keep the word 'forum', but get rid of the word 'megascience'," says Michael Osborne, associate director of the OECD's science, technology and industry directorate.

When the Megascience Forum was set up in 1992 there was intense interest in the potential rivalry between the US Superconducting Collider and Europe's proposed Large Hadron Collider. But the demise of the US facility, and the apparent waning of political interest in other 'big science' projects in areas such as space and fusion research, cast doubt on the forum's continuing validity.

More recently, however, the forum has

successfully shifted its focus to topics of more immediate interest to agencies that sponsor science, such as the terms of access to large-scale facilities and the availability of neutron sources (see box).

Member states are also said to have valued highly the discussions of a working group that led to the proposal to establish a Global Biodiversity Information Facility linking biodiversity databases around the world (see *Nature* **394**, 118; 1998).

Those involved in the biodiversity database proposal say it is the type of activity in which the forum has proved its potential value in bringing together scientists and policy-makers, even though it does not qualify as big science.

Another study, on radioastronomy, played a key role enabling researchers and representatives of the telecommunications industry to discuss potential threats to astronomical observations from portable telephones and other devices.

Osborne emphasizes that the new forum, if given formal approval, will ensure that the topics it studies remain close to the immediate concerns of member governments. "The forum has to be action-oriented," he says. "Governments must not be just concerned about an issue, but under some pressure to do something about it."

One possibility is that the forum will look at techniques for nuclear waste disposal, a hot potato in many parts of the world, particularly now that a failure to resolve the issues thrown up by scientific debates on the safety of disposal techniques threatens what many see as a major potential answer to the problem of global warming.

"We do not want to be just another bunch of bureaucrats complaining that science does not have enough money," says Stefan Michalowski, the forum's executive secretary. "We want to have government people, backed up by scientists, trying to figure out how to pool resources and how to generate material that is useful to them."

Michalowski also points out that the forum allows Japan, as a member of the OECD, to take an active part in such debates (a senior official of Japan's Science and Technology Agency, is its vice-chair). "If Megascience has anything going for it, it is the way it provides a bridge between Japan and other OECD countries."

The final structure of the new body, as well as its precise terms of reference, are yet to be agreed, and such decisions will be taken in the context of a broader reform of the OECD's science activities. But with key member states recognizing the value of a forum at which senior science policy-makers can get together, its immediate future seems assured.

David Dickson

...and warns of need for more neutron sources

[PARIS] The 6,000-strong international community of users of neutron scattering techniques needs to take a longer-term view of future needs for large machines, and be better organized in mustering political support, if it is to enjoy an adequate supply of neutrons early in the next century.

This is a key message of a report produced by the Megascience Forum of the Organization for Economic Cooperation and Development (OECD). It predicts that the world supply of neutron sources will fall far short of future demand early next century unless decisions to build new ones are taken within the next five years*.

There are now around 26 major neutron sources in the OECD countries and Russia. But most will reach the end of their lifetimes between 2005 and 2015, and the report warns that, unless prompt action is taken, the capacity of neutron sources in 2010–20 will be one-third of that available today.

Andrew Taylor, director of the ISIS neutron facility in the United Kingdom, endorses the report's comment that, whereas particle physicists plan ahead for the large facilities on which they

depend, neutron users have been less well organized.

The problem, says Taylor, is that neutron sources are not big machines used by a few, but a shared tool used by researchers from many disciplines, often for short periods. "It has been difficult to get young scientists interested in something that is not going to exist for twenty years," he says.

The report warns that a neutron drought would be a "serious threat" to scientists working in many disciplines, including biology, Earth sciences, engineering and materials science. At present, about 4,000 researchers in Europe use neutron scattering techniques, with another 1,000 in the United States and 1,000 in Japan.

The greater activity in Europe reflects the presence of the world's two most powerful neutron sources, the Institut Laue-Langevin (ILL) in France and ISIS. The United States, which was at the forefront of neutron scattering in the 1960s, has since fallen behind.

But the report points out that the planned US\$1.3 billion national Spallation Neutron Source at Oak Ridge, Tennessee, will allow the United States to catch up with Europe. The project was

approved in the US budget last month (see *Nature* **395**, 531; 1998). Japan has plans for similar facilities.

Europe is pressing ahead with plans for a new five-megawatt European Spallation Source, now in the research and development phase. Its proponents hope to complete this by 2003, and to get political backing by 2005 for the machine to come on-line by 2016.

The report encourages neutron users to cooperate in putting together detailed plans for new machines and attracting political support, and says that greater attention needs to be paid to financing adequate instrumentation.

It points out that the ILL reactor provides neutrons to 43 instruments, used by about 1,200 scientists a year from disciplines such as biology chemistry and solid-state physics, whereas the similarly sized source High-Flux Isotope Reactor at Oak Ridge, has only 11 instruments. "New sources attract bright young people and lead to development of new instruments," adds Taylor.

Declan Butler

* Richter, D. & Springer, T. A *Twenty Years Forward Look at Neutron Scattering Facilities in the OECD Countries and Russia*. Available on <http://www.oecd.org/dsti/sti/>

Industrial links still face opposition in Japan's universities

Asako Saegusa

Japan has traditionally had poor links between universities and industry. The government hopes to change this. But its efforts face a daunting challenge in the need to transform deeply engrained attitudes.

[TOKYO] Japan's Ministry of International Trade and Industry (MITI) announced last week that it is setting up a ¥50 billion (US\$427 million) fund to support up to 1,000 'venture businesses' a year, including some based at universities. The move is part of the government's effort to promote the commercial application of university research by improving links between universities and industry.

But substantial changes in attitude are required for such endeavours to succeed. For example, the current law on civil servants forbids researchers at national universities from getting involved with profit-making activities and carrying out research at industrial laboratories.

In addition, Japanese universities have a conservative attitude towards such activities. Researchers point out that making inventions and obtaining patents do not earn them any credit in institutions that still see the production of scientific papers as their top priority.

Since a law promoting collaboration between universities and industry was passed in May, both the Ministry of Education, Science, Sports and Culture (Monbusho) and MITI have actively backed programmes to set up technology liaison offices (TLOs) and venture businesses at Japanese universities (see *Nature* 390, 105; 1997).

TLOs have already been established at some national universities, including Tokyo University and Tsukuba University. They are intended to help draw up contracts for collaboration with industrial laboratories and to advise on the commercial exploitation of inventions made by university researchers.

Monbusho has also been backing Venture Business Laboratories (VBLs), a programme begun in 1996 to develop closer ties between academic research and the industrial community. VBLs have so far been set up at 24 national universities, providing laboratories for joint research projects and training programmes for students and researchers who wish to set up their own small companies to commercialize university research.

The government has already started changing the regulations on university fund-



Venture capital: Tsukuba University is investing in new companies that find uses for research.

ing to encourage such developments. For example, Monbusho has relaxed its restrictions on universities applying for funds from other government sources, such as MITI. Another change allows private investors to lend money to universities creating start-up companies based on their scientific discoveries; private funding to national universities was previously restricted to research grants, scholarships and donations.

But restrictions on university staff mean that new ventures, such as the Centre for Advanced Science and Technology Incubation (CASTI), a limited holding company set up in July by professors from Tokyo University's Research Centre for Advanced Science and Technology (RCAST), have had to recruit staff from outside the university.

CASTI is intended to evaluate the marketability of inventions by university researchers, draft patent applications, and sell patent rights to private companies (see *Nature* 391, 622; 1998). It is the first unit set up by a Japanese university to promote the commercial exploitation of its inventions. Some of the royalties from sales of the patents will go to the inventor, with the rest being paid into RCAST's research budget.

Because RCAST belongs to a university, it will not be able to take an active role in the business or hold patent rights to inventions made at the institute, and the professors who set up RCAST and control its work are not allowed to receive any profits, as this would infringe the civil-service law.

The conservatism of Japanese universities is reflected in the fact that they produce very few patents. In 1994, for example, only

124 out of 351,300 patents filed internationally were from universities, and just seven out of 215,400 domestic patents filed in 1996 were from Tokyo University.

According to a recent survey by *Nihon Keizai Shinbun* (*Nikkei*), a daily financial newspaper, only 30 per cent of universities surveyed expressed support for their academic staff's involvement in business activities. "Given the current legal restrictions, national universities are still wary about getting involved in business activities," says a professor from RCAST. "There is an inevitable conflict between the 'intellectual community', which strongly believes that universities should have nothing to do with profit-making activities, and business-minded academics, who are keen to change such an attitude."

The survey also showed that most universities are reluctant to provide financial support for new venture businesses, and more than two thirds said they do not intend to create new funds to support such activities.

An exception to this is Tsukuba University. Together with JAFCO, Japan's largest venture-capital company, it has set up a ¥1 billion fund to invest in venture businesses.

Tamon Inoue, professor in engineering at Tsukuba University and head of its VBL, says it will not be easy to break the conservative attitude of universities towards the creation of new businesses, or to encourage students to become entrepreneurs.

Furthermore, with the current economic turmoil, small high-tech companies are struggling to survive, as banks and private investors are reluctant to lend and invest money. "University students still hold firmly to the idea of lifetime employment, and joining or setting up venture businesses seems to be out of their scope," says Inoue.

He also points out that research at universities rarely lends itself to commercial application. "The focus of industrial research and university research are fundamentally different," he says. "For businesses to flourish from universities, researchers will need carry out 'useful' research that could bring profit to the university."

However, while researchers keen to see more collaboration between universities and industry are calling for a more flexible system that gives them more freedom, not all support the government's proposal to transform such universities into semi-autonomous institutions (see *Nature* 395, 730; 1998).

They accept that more autonomy would provide better opportunities for new arrangements with industry to thrive. But it is uncertain whether the new system would free university researchers from civil-service rules. Many researchers say the reforms should be carried out slowly, as a quick transformation could throw the universities' management systems into turmoil. □

TSUKUBA UNIVERSITY

WHO's anti-malaria campaign wins interagency backing

[PARIS] Plans by Gro Harlem Brundtland, the new director general of the World Health Organization (WHO), to make fighting malaria a priority of her term in office moved forward last week with an agreement to create a multi-agency programme for malaria research and control.

The Roll Back Malaria campaign brings together the WHO, the World Bank, the United Nations Children's Fund and the United Nations Development Programme. All have agreed to incorporate malaria initiatives across their programmes.

Under the agreement, signed in New York, the WHO pledged to boost research networks and programmes, while the World Bank said it would increase its investment in malaria research and control, and explore novel financing mechanisms for drugs research and supply (see *Nature* 395, 417–418; 1998).

The initiative has the backing of prominent research figures such as Harold Varmus, director of the US National Institutes of Health, and Fotis Kafatos, director general of the European Molecular Biology Laboratory in Heidelberg, Germany.

It will be headed by David Nabarro, formerly director of health and population at the UK Department for International Development.

US industrial research still set to climb next year

[WASHINGTON] **Industrial support for research and development (R&D) in the United States is expected to keep growing next year, but at a slower rate than this year, according to a survey by the Industrial Research Institute of 106 member companies, including many of the biggest research performers in US industry.**

The annual survey found that 22 per cent of research managers in large corporations expected to boost their R&D spending by more than 5 per cent in the coming year; last year, 31 per cent expected to do this. The institute says that recent worries about a pending global economic crisis have resulted in only a modest downward correction in corporations' plans to invest in research.

India brings in welfare controls for lab animals

[NEW DELHI] The Indian government announced last week that all recognized scientific institutions seeking to import animals for research must obtain permission

from a committee set up by the ministry of environment and forests that is concerned with preventing cruelty to animals.

The committee is headed by Maneka Gandhi, India's welfare minister and an animal activist who believes that too many animals are killed in schools and labs. The move comes despite warnings from the heads of the Indian Council of Medical Research and the Indian National Science Academy that it could hamper drug research.

Coral reef damage acts as a climate warning

[SYDNEY] **Climate change treaty negotiators should take account of reef ecology as a sensitive indicator of global warming, suggests the International Society for Reef Studies. The researchers say this factor was omitted from the agenda at last year's climate conference in Kyoto.**

The 750-member society issued a statement aimed at those attending the meeting on the United Nations climate change convention which begins this week in Buenos Aires (see page 7). It says coral bleaching has been found in at least 32 countries and island nations during the past year — "the most geographically widespread bleaching ever recorded". This is "a major cause for concern," concludes the statement.

Labs to terminate use of 'terminator' gene

[WASHINGTON] The Consultative Group on International Agricultural Research (CGIAR), which oversees the \$350 million agricultural research programme carried out in developing countries by the World Bank and several United Nations agencies, has banned 'terminator' plant genes from its laboratories and research stations.

Terminator genes, which ensure that plants do not produce their own seed, forcing farmers to repurchase it each year, are being developed by several agricultural biotechnology corporations to protect their investment in other crop modifications.

But the CGIAR, meeting in Washington, agreed not to use "any genetic system which is designed to prevent seed germination", saying that such systems were damaging to the interests of poor farmers, bad for genetic diversity, and pose a risk to other plants with which crops may cross-pollinate.

Nanjing tops the citation charts in China again

[TOKYO] Nanjing University in southern China has again come top in China in output of scientific papers in the Science Citation Index (SCI), ahead of the two leading

universities in the capital, Beijing. In 1997, Nanjing produced 682 papers listed in the index, 112 up on 1996. Peking University came second, with 448 papers, and Tsinghua University had 407.

The figures were released two weeks ago by China's Ministry of Science and Technology. But China's universities are still far behind leading universities in developed nations which typically produce thousands of SCI papers a year.

Leukaemia scare shuts Mozarteum building

[SALZBURG] The main building of the Mozarteum, Salzburg's world-renowned academy of music, closed last month after high levels of airborne carcinogens were detected in the 20-year-old building.

The Salzburg health office commissioned an investigation after three professors and one student had contracted leukaemia in recent years, and many other teachers and students had complained of serious health problems.

As yet no causal link between the carcinogens detected and the leukaemia cases has been proven. Last week the Salzburg authorities arranged blood screening of 20 long-serving Mozarteum teachers, and free medical and psychological

care for all the staff and students. A committee of experts has been established to try to determine the source of the contamination.

Looking to history for a more modern image



[LONDON] The Royal Society, Britain's 338-year-old academy of science, has adopted a new logo (above) following a review of how it communicates with the outside world. The logo's letters, 'Sc', although conveying a modern appearance, are based on the first serif typeface carved in the second century AD on Trajan's column in Rome.

The review was carried out by Synergy Communications, which concluded that the society needed a contemporary identity that also communicated its long tradition of excellence in science.

The society reveals in its latest newsletter the challenge of conveying its identity without being able to use the word 'science' in its title. "A perennial question has been 'The Royal Society... of what?' The new identity should help towards making that query redundant."

Generation of young scientists in peril

Sir—As a former scientist married to someone who is in his second year of a postdoctoral fellowship, I could not agree more with John Moore's comments on the sad state of funding for young scientists in the United States (*Nature* **395**, 431; 1998). I suspect that the career situation for scientists will get worse before it gets better, resulting in the frustration and demoralization of a generation of bright young people. I hope that the voices of people such as Moore will be heard in Congress before the next generation of scientists is lost.

My decision to leave my postdoc was based on several factors, but one of the primary reasons was concern that we would not be able to support a family on two postdoc salaries. Most of my colleagues believed that I would be able to succeed in academic research, but they all supported my decision then and agree now that it was the right move. Although my husband is

committed to research, we sometimes wonder about our financial future.

I now work at a public relations firm that specializes in the biotechnology, pharmaceutical and healthcare sectors. In addition to providing better wages and benefits, this position offers something that is sorely lacking in academia: recognition and reward for a job done well. I have twice been promoted, and get daily positive reinforcement from co-workers and clients.

My office recently hired a friend of mine from graduate school, my sister left science for a job on Wall Street and another friend is thinking of leaving her postdoc for a job in industry where she could earn more and have more time to spend with her child. Judging from the résumés I receive each month from scientists seeking new job opportunities, we are not alone.

Without changes to the current system, I fear that financial realities will force a growing number of talented people to seek

alternative careers that will enable them to use their hard-earned degrees for greater monetary reward.

Stephanie Seiler

Noonan/Russo Communications,
220 Fifth Avenue, New York 10001, USA

Sir—John Moore rightly suggests that one way to attract bright US students to postdoctoral positions, and eventually to academia, is to raise their salaries. But his concluding sentence appears to suggest the wrong solution.

If the government spends money to "train more young scientists" this should tend to create a further glut of postdocs, and depress their salaries further. To raise salaries, we should be funding fewer young scientists in order to pay them individually more with the same pot of money.

Keith Alverson

Past Global Changes International Project Office,
Bern, Switzerland

Wronged by Crookes

Sir—The 10 September Daedalus column was, as usual, fascinating and charming (*Nature* **395**, 120; 1998). But it raises the possibility of a grave academic injustice dating from nearly 20 years ago.

At the time, I was a sophomore in a prestigious university in the American Midwest (we'll leave out the institution's name, but its initials are University of Chicago). I recall with chagrin breaking up on the rocks of an introductory quantum mechanics examination (I had an A grade going into the final) in which one of the questions was an explanation of the Crookes radiometer, as dealt with by Daedalus.

Never mind what I answered; you're not going to get me to admit to that. But the correct answer, I was told by a tired-looking professor, was that photons hitting the black side of the vane were experiencing an inelastic collision, while those hitting the white side were colliding elastically. The inelastic collision imparted a momentum, P (derived from the photon), onto the vane, but in order to conserve momentum the elastic one had to impart a $2P$ momentum. Hence, motion in the direction of the black side of the vanes, QED.

I existed at the time in a perpetual state of indignation, which was not helped by the intrusion of classical mechanics into my world of hamiltonian operators and eigenfunctions (any misuse of terms, by the way, I blame on my teachers).

Now, however, Daedalus provides a

mechanism of action for the Crookes radiometer that deviates from the given wisdom of that fateful day, nigh on two decades ago. Have I been wronged? Should I bring this case before the university's ethics committee? I am, of course, willing to dispense with any considerations of academic collegiality or respect for a talented researcher/teacher. I'm a journalist now, I need neither.

I await a reply with high expectations of twisting that A out of said professor. Better late than never.

Kenneth B. Chiacchia

Public Relations, UPMC Health System, 200 Lothrop
Street, Pittsburgh, Pennsylvania 15213, USA

On the world stage

Sir—Your article "Urgent thinking required about development" refers to the World Conference on Science to be organized next year in Budapest by the United Nations Educational, Scientific and Cultural Organization (*Nature* **395**, 527; 1998). In fact this conference is jointly sponsored by UNESCO and the International Council for Science (ICSU).

This is the first time that UNESCO, a governmental organization, has joined with a non-governmental organization to co-sponsor such an international event.

The conference will be held from 26 June to 1 July, not in May as stated in your article, and will deal with such topics as sustainable development, science and

development, science and industry, and knowledge as a public good.

Jean-François Stuyck-Taillandier

(Executive director)

ICSU, 51 Bld de Montmorency, 75016 Paris, France

Fishing for compliments

Sir—There is an alternative explanation for the cause of citation errors to that offered by Nicholas Price (*Nature* **395**, 538; 1998).

In 1957, I was the author, with R. J. H. Beverton, of a book (*On the Dynamics of Exploited Fish Populations*, HMSO) which, according to citation indexes, is still the most cited reference in fisheries science. We both occasionally wondered why that should be when the book was long out of print—until republished in 1993—and not readily available outside specialized libraries. It seemed that it became 'politically correct' for newcomers to our field to cite our work, and a considerable proportion of them had clearly never read or even seen it, as we could tell from internal evidence in their publications.

So one or two incorrect references reproduced themselves; they were not mistakes or typos or due to simple carelessness. Naturally, we were proud of our citation records, but, as we say here, it is *un po' esagerata*.

Sidney Holt

Podere il Falco, 06060 Ponticelli PG, Italy

Founding father

Eric S. Lander and Joseph J. Ellis

Almost two hundred years ago Thomas Jefferson was alleged to have fathered a children by his slave Sally Hemings. The charges have remained controversial. Now, DNA analysis confirms that Jefferson was indeed the father of at least one of Hemings' children.

For two centuries Thomas Jefferson's legacy has been haunted by the first US presidential sex scandal — the charge of an illicit relationship with his mulatto slave Sally Hemings. From the day the story broke in a Richmond newspaper in 1802, 'Tom and Sally' has become the longest running mini-series in American history. Because the evidence was all circumstantial, no authoritative resolution has been possible. Until today, that is. On page 27 of this issue, Foster *et al.*¹ report that DNA testing of Y chromosomes offers strong evidence that Jefferson fathered at least one of Hemings' children.

The saga begins in the mid-1780s in Paris, where Jefferson served as ambassador to France after the death of his wife. Sally Hemings, then 14 years old, was sent to accompany Jefferson's youngest daughter to Paris in 1786. There is no evidence of what transpired there, but Hemings returned to the United States with Jefferson in 1789, and she eventually bore at least five children, starting with Tom in 1790 and ending with Eston in 1808.

At least three pieces of evidence support a relationship between Jefferson and Hemings. First, several of the children bore a

striking physical resemblance to Jefferson. Second, Sally's fourth child, Madison, testified late in life that Sally had identified Jefferson as the father of all her children. Finally, Jefferson was in residence at his mansion in Monticello in Virginia at the time when each of the children was conceived. But many historians have expressed doubts, and Jefferson family tradition has implicated a maternal cousin as the likely father.

To a geneticist, the obvious solution — short of exhuming the principals — is to compare Y chromosomes from modern-day male-line descendants. Most of the Y chromosome is passed intact from father to son, so it can be used to trace paternal lineages. However, such studies require enough polymorphic markers (small regions of DNA that vary among individuals) so that Y chromosomes can be distinguished by the haplotype (set of specific variants) that they carry. Researchers from several laboratories have identified a collection of suitable markers from the Y chromosome over the past two years, and this collection is now fuelling an explosion in male-line genetic studies.

Foster *et al.*¹ examined a haplotype containing 19 polymorphic markers. Jefferson's haplotype (inferred from male-line descen-

Name	Date	Other
Ned	5/2	7
Jerry	5	7
Moses	3	4
Suskey	6	3/4
Nedj	5/2	7
Peter Hem	5/2	7
Philip	5/2	7
Rachael	5	7
Eliza	5	3/4
Ellen	8	3/4
Sally	5	7
Harriet	1	4/5
Madison	2	3/4
Eston	8	3/4
Samed	5/2	7

Figure 2 A portion of Jefferson's farm book, in which he recorded the distribution of slave rations by family.

dants of his paternal grandfather) seems to be quite rare, inasmuch as it was not seen among a sample of 670 Europeans or 1,200 people worldwide. The authors found that this rare haplotype perfectly matches that of Eston Hemings' male-line descendant. The probability of such a match arising by chance is low — safely less than 1%. Together with the circumstantial evidence, it seems to seal the case that Jefferson was Eston Hemings' father.

Interestingly, Jefferson's haplotype does not match male descendants of Sally's first son, Tom Woodson. The simplest explanation is that Jefferson was not Tom's father. An alternative explanation would require non-paternities among Tom's offspring. The jury remains out with respect to Sally's other children, but the burden of proof has clearly shifted.

Nothing in Foster and colleagues' study, and nothing in the vast historical literature, sheds any light on the character of the relationship between Jefferson and Sally Hemings. Was it, as his contemporary critics charged, a tale of lust and rape? Was it, as several twentieth-century scholars and novelists have suggested, a love story rooted in mutual affection? Or was it something in-between? These questions are open to endless interpretation but, in a broader sense, the new findings give blacks and whites alike an opportunity to confront a largely secret, shared history.

Politically, the Thomas Jefferson verdict is likely to figure in upcoming impeachment hearings on William Jefferson Clinton's sexual indiscretions, in which DNA testing has also played a role. The parallels are hardly perfect, but some are striking. Both 'improper' relationships involved women about 28 years younger — although there is a world of difference between a slave and master at the close of the eighteenth century, and a White House intern and a married man at the end

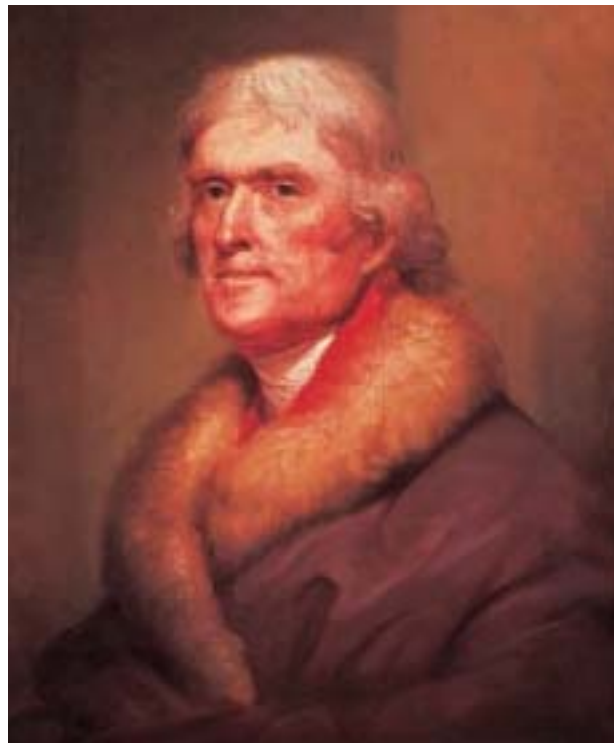


Figure 1 Thomas Jefferson (1743–1826), third president of the United States. DNA analysis by Foster *et al.*¹ shows that he fathered at least one child by his slave, Sally Hemings.

of the twentieth. Both presidents seem to have engaged in politically reckless conduct; in Jefferson's case, fathering Eston six years after allegations appeared in the national press. And both offered evasive denials to the charges. In 1805 the Massachusetts legislature staged a mock impeachment trial of Jefferson, citing several grievances including the accusations about Sally Hemings. Jefferson acknowledged one charge (propositioning a married woman in his youth), but asserted that all the others were false. Otherwise he remained silent, leaving denials to political supporters and family. Nor did the scandal affect Jefferson's popularity. He won the 1804 election by a landslide, and his abiding position was that his private life was nobody else's business, and should have no bearing on his public reputation.

Foster and colleagues' findings renew questions about Jefferson's tortured position on slavery. If Jefferson's relationship with Hemings began in the late 1780s, it would mean that he began to back away from a leadership position in the anti-slavery movement just around the time that his affair with Sally Hemings started. Jefferson's stated reservations about ending slavery included a fear that emancipation would lead to racial mixing and amalgamation. His own interracial affair now personalizes this issue, while adding a dimension of hypocrisy.

Over the past 30 years, research into Jefferson has cast a shadow over his credibility as America's prophet of freedom and equality. Recent work has also emphasized his massive personal contradictions and his dexterity at playing hide-and-seek within himself. The new evidence only deepens the paradoxes.

Jefferson is, with Abraham Lincoln and George Washington, one of America's secular saints. His face looks out from the nickel, the two-dollar bill, the memorial near the Tidal Basin, and Mount Rushmore. His unique capacity to project inspirational words and ideas onto American public life has made him all things to all people. As an icon, Jefferson's legacy has been reinterpreted by every generation. Now, with impeccable timing, Jefferson reappears to remind us of a truth that should be self evident. Our heroes — and especially presidents — are not gods or saints, but flesh-and-blood humans, with all of the frailties and imperfections that this entails. □

Eric S. Lander is at the Whitehead Institute for Biomedical Research and the Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA.

e-mail: lander@genome.wi.mit.edu

Joseph J. Ellis is in the Department of History, Mount Holyoke College, South Hadley, Massachusetts 01075, USA.

I. Foster, E. A. et al. *Nature* **396**, 27–28 (1998).



Figure 1 Penrose's planar quasiperiodic tiling. It is formed by a set of two rhombuses with edges of equal length, one with angles of 36° and 144° and one with angles of 72° and 108°. Their edges are marked with single or double arrows, and the rules constrain adjacent tiles to have matching arrow types along their shared edge.

adopt 'vertex' rules⁶, which can be interpreted in terms of short-range interactions between atoms in clusters centred on a given vertex. But this mathematical exercise is of little use to the physicist who wants to understand why and how quasicrystals form, as the tiles correspond neither to atoms nor to real atom clusters.

In 1991, Sergei Burkov⁷ realized that planar quasiperiodic tilings can be generated with only a single tile, a decagon, provided that the tiles can overlap. Five years later, Petra Gummelt⁸ gave a mathematical proof that a quasiperiodic Penrose tiling can be generated using a single decagon combined with a novel overlapping rule. This rule is realized by decorating the interior of the decagons (Fig. 2) with a subset of shaded tiles, and two decagons may overlap only if shaded areas overlap. This is equivalent to Penrose's arrow matching rules⁹. In tilings overlapping has to be avoided as a point of principle. Here it becomes a basic construction element; so, using an established mathematical term¹⁰, Gummelt called the pattern a 'coverage'.

Hyeong-Chai Jeong and Steinhardt⁹ then proved that Penrose matching rules can be abandoned altogether and replaced by the condition that the density of a suitably chosen cluster (for instance in the form of Gummelt's decagon) is maximized. And this is where mathematics at last leads to physics. Jeong and Steinhardt concluded that quasicrystals represent a packing of a single type of atom cluster. This cluster can share atoms with its neighbours, and the resulting quasiperiodic pattern is just the one that maximizes cluster density. By postulating that this cluster corresponds to a minimum-energy atom configuration, the authors arrived at a physically plausible picture.

Striking evidence for the coverage model

Quasicrystals

From tilings to coverings

Knut W. Urban

Quasicrystals occur in a great number of alloys, most of which consist of aluminium and transition metals. They were discovered¹ in 1984, being revealed by a rotational symmetry of X-ray or electron diffraction patterns (for instance, five-fold or ten-fold) which is impossible for true periodic crystals. Since then, physicists have wondered why atoms form in these complex patterns rather than in a regularly repeating periodic crystal. On page 55 of this issue² Paul Steinhardt and colleagues present an analysis of new electron microscope data that supports a simple answer to this question.

Although a quasicrystal is non-periodic, its structure still follows a subtle construction plan. Mathematically, this can be described with reference to a higher-dimensional analogue of a cubic lattice: the atom arrangement of an icosahedral quasicrystal (which is quasiperiodic in three dimensions) can be constructed starting from six-dimensional space; the decagonal quasicrystal (whose lattice is quasiperiodic in a plane but periodic along the third dimension) requires reference to five-dimensional space. But why should atoms care about higher-dimension-

al spaces? The mathematical recipe to describe the lattice does not give us any hint as to how the atoms manage to create it.

Ten years before their discovery in nature, quasiperiodic patterns with the same geometric properties as those calculated from five-dimensional hyperspace were described by Roger Penrose³. These are tilings of the plane, in which a set of suitable tiles is arranged without gaps or overlaps according to certain matching rules. An example (Fig. 1) is the set consisting of two rhombuses with edges of equal length, one with angles of 36° and 144° and the other with angles of 72° and 108°. Their edges are marked with single or double arrows, and the rules constrain adjacent tiles to have matching arrow types along their shared edge. Corresponding matching rules have been found for three-dimensionally quasiperiodic patterns based on two types of rhombohedron⁴.

It would appear that Penrose's 'edge' rules can be used to mimic growth of quasicrystals by stepwise addition of tiles to a seed. However, because these rules are strictly local in nature, they are not enough to guarantee a defect-free quasiperiodic pattern⁵. A solution to this growth problem is instead to

now comes from electron microscopy on decagonal $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$. Steinhardt and colleagues² present micrographs that demonstrate the existence of Gummelt's decagonal coverage in nature. They used the high-angle annular dark-field technique¹¹. In contrast to classical high-resolution transmission electron microscopy, where contrast interpretation requires a careful quantum-mechanical simulation of the dynamic imaging process (often with rather ambiguous results), the dark-field technique is based on high-angle scattering of a fine beam of electrons at atom cores in a scanning transmission electron microscope. In this case the imaging process is essentially kinematic in nature and allows a direct, intuitive image interpretation in terms of atomic structure. This makes the new results superior to those of previous studies in which atomic matching of classical high-resolution images to model lattices has been claimed.

Steinhardt and colleagues call the decorated decagon a 'quasi-unit cell' to emphasize similarities to ordinary crystal lattices. Whether this is a happy choice is questionable as clusters often also play a role in crystals; an example is the Frank–Kasper phases (where clusters occur which are quite similar to those found in quasicrystals). But these clusters never coincide with the crystal unit cells.

That clusters are important for quasicrystal formation was noticed quite early on^{12,13}. In particular, the model of icosahedral Al-Pd-Mn , the quasicrystal of highest structural quality available today, is based on a

52-atom icosahedral cluster¹⁴. In this case it could even be shown that it owes its particular stability to a low electron energy¹⁵. In its present form, however, this model does not explicitly involve cluster overlap as a construction principle, but employs 'connecting units' which are interpenetrating pieces of the elementary cluster. Indeed, the new 'covering' model has still to be tested for three-dimensionally quasiperiodic quasicrystals. But the idea that the particular structure of quasicrystals is the consequence of maximizing the density of a low-energy atom cluster and (what appears to be particularly important) the experimental observation of the Gummelt coverage, make it worth while revising current models to incorporate this new view of quasicrystal formation. □

Knut W. Urban is at the Institute of Solid State Physics, Helmholtz Research Centre Jülich,

D-52425 Jülich, Germany.

e-mail: k.urban@fz-juelich.de

1. Shechtman, D., Blech, I., Gratias, D. & Cahn, J. W. *Phys. Rev. Lett.* **53**, 1951–1953 (1984).
2. Steinhardt, P. J. *et al.* *Nature* **396**, 55–57 (1998).
3. Penrose, R. *Bull. Inst. Math. Appl.* **10**, 266–271 (1974).
4. Katz, A. in *Quasicrystalline Materials* (eds Janot, C. & Dubois, J. M.) 195–204 (World Scientific, Singapore, 1988).
5. Van Ophuyssens, G., Weber, M. & Danzer, L. *J. Phys. A* **28**, 281–292 (1995).
6. Onoda, G. Y., Steinhardt, P. J., DiVicenzo, D. P. & Socolar, J. E. S. *Phys. Rev. Lett.* **60**, 2653–2656 (1988).
7. Burkov, S. E. *Phys. Rev. Lett.* **67**, 614–617 (1991).
8. Gummelt, P. *Geometriae Dedicata* **62**, 1–17 (1996).
9. Jeong, H.-C. & Steinhardt, P. J. *Phys. Rev. B* **55**, 3520–3532 (1997).
10. Grünbaum, B. & Shephard, G. C. *Tilings and Patterns* (Freeman, New York, 1987).
11. Pennycook, S. J., Jesson, D. E., Nellist, P. D., Chrisholm, M. F. & Browning, N. D. in *Handbook of Microscopy Vol. II* (eds Amelinckx, S. *et al.*) 595–620 (Wiley-VCH, Weinheim, 1997).
12. Guyot, P. & Audier, M. *Phil. Mag.* **B52**, L15–L19 (1985).
13. Elser, V. & Henley, C. L. *Phys. Rev. Lett.* **55**, 2883–2886 (1985).
14. Boudard, M. *et al.* *J. Phys. Condens. Matter* **4**, 10149–10168 (1992).
15. Janot, C. *Phys. Rev. B* **53**, 181–189 (1996).

Inflammation

A new target for aspirin

Edward A. O'Neill

"Take two aspirin and call me in the morning" seems like a tired cliché. Nevertheless, for many common physical ailments, low doses of aspirin (acetylsalicylic acid) will make you feel better, and high doses of the drug give sustained relief from the symptoms of chronic inflammatory diseases such as rheumatoid arthritis¹. Aspirin irreversibly inhibits the cyclooxygenases — enzymes that control the production of prostaglandins (small molecules that induce the pain and fever associated with infection or trauma²). However, the difference in the clinical activities of aspirin at low and high doses has led to speculation that not all the benefits of aspirin derive from inhibition of cyclooxygenases. So what might be the target for high doses of the drug? On page 77 of this issue, Yin *et al.*³ report that high concentrations of aspirin inhibit the recently discovered enzyme I κ B kinase- β (IKK- β), and they propose that this effect partially explains the clinical efficacy of high-dose aspirin.

The IKK enzymes (α and β) catalyse the transfer of phosphate moieties from ATP to I κ B (reviewed in ref. 4). Phosphorylation leads to the degradation of I κ B and release of NF- κ B, a transcription factor that is inhibited by I κ B. On release, NF- κ B rapidly moves to the nucleus where it binds specific DNA sequences, promoting the transcription of genes that influence defence mechanisms such as inflammatory and immune responses. Thus, if IKK- β is inhibited by aspirin, nuclear localization of NF- κ B and subsequent transcription should be blocked. This has, in fact, been observed with high concentrations of aspirin⁵.

Yin *et al.*³ now report that high concen-

trations of aspirin (IC_{50} approximately 50 μM) are required to inhibit IKK- β . They also show that several other kinases, including the homologous and functionally related IKK- α , are not affected by aspirin. Can inhibition of IKK- β and cyclooxygenases explain all the pharmacological effects of aspirin? Let's hope not. The problem with using high doses of aspirin is that it has virtually no therapeutic window — that is, the dose at which aspirin gives relief from chronic rheumatic disease is very close to the dose that generates side-effects, including headaches, dizziness and tinnitus⁶. But, based on the work of Yin and colleagues, we can develop a hypothesis wherein IKK- β is only one of several targets for high-dose aspirin, the unwanted side-effects resulting from activity against other targets.

The anti-inflammatory efficacy of aspirin may, therefore, be limited by two separable factors. First, aspirin weakly inhibits the therapeutic target (IKK- β). Second, it is a similarly weak inhibitor of other, as-yet-unidentified targets, yielding various toxicities. There are many reports of aspirin affecting a variety of biological systems that could be independent of IKK- β inhibition. Most interesting in light of the paper by Yin *et al.* are reports that high concentrations of aspirin inhibit the activation of the Jun-N-terminal kinases (JNKs)⁷. The JNK family of protein kinases mediates phosphorylation of the transcription factor c-Jun, stimulating its ability to promote transcription⁸. Because there is no evidence that IKK- β is involved in activating JNK, these results imply that at least one other target of high-dose aspirin exists.

Although there is a broad consensus that aspirin relieves many of the symptoms asso-

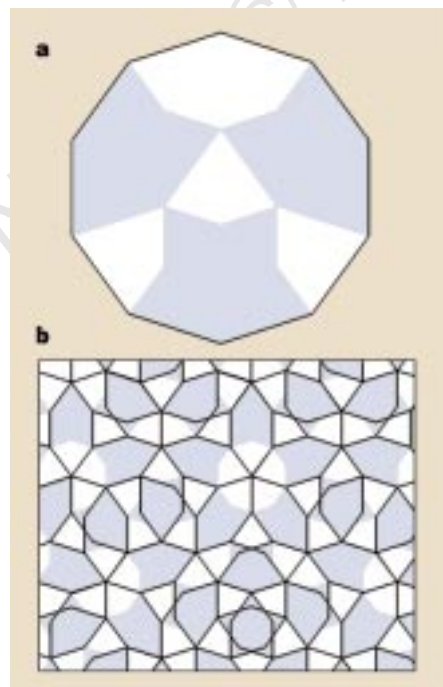


Figure 2 Gummelt's planar quasiperiodic coverage. a, Gummelt's decorated decagon. The matching rule demands that two decagons may overlap only if shaded areas overlap. b, The resulting quasiperiodic decagonal coverage of the plane.

ciated with inflammatory diseases, it does not prevent the hallmark of progressively crippling rheumatoid arthritis — the destruction of joints. This could be due to the limited effect that aspirin can have on NF- κ B activation, constrained as it is by the narrow therapeutic window. The most effective drugs for reducing joint destruction are immunosuppressants such as prednisolone and cyclosporin A. These drugs work because they are extremely effective at inhibiting the transcription of genes that encode pro-inflammatory cytokines. Unfortunately, both have dose-limiting toxicities associated with the mechanisms by which they function. Cyclosporin A inhibits calcineurin, an enzyme that regulates the activity of transcription factors (therapeutic effect) in some cells, but also regulates the function of ion channels (toxic effect) in other cells. Prednisolone binds to the glucocorticoid receptor, which then represses transcription (therapeutic effect) in some cellular contexts, but activates transcription (toxic effect) in others. There is limited hope of separating the undesirable from the desirable effects of these drugs.

On the other hand, thanks to the work of Yin and colleagues, we can consider two new approaches to separating the therapeutic from the toxic effects of high-dose aspirin.

First, IKK- β might be an excellent molecular target for anti-inflammatory drugs, giving pharmacologists a simple, quantitative assay for the therapeutic activity of new chemical entities. Second, we now know that selective, but weak, inhibitors of IKK- β exist, giving medicinal chemists a starting point from which to build more potent inhibitors with greater target selectivity. Using these tools, we may be able to generate small molecules that can prevent rheumatoid joint destruction, rather than just relieving the symptoms associated with that destruction. Until then, take two aspirin and call me in the morning. □

Edward A. O'Neill is in the Department of Immunology and Rheumatology, Merck Research Laboratories, PO Box 2000, Rahway, New Jersey 07065-0900, USA.

e-mail: edward_oneill@merck.com

1. Hart, F. D., Huskisson, E. C. & Ansell, B. M. in *Drug Treatment of the Rheumatic Diseases* (ed. Hart, F. D.) 7–60 (ADIS, New York, 1982).
2. Vane, J. R. *Nature New Biol.* **231**, 232–235 (1971).
3. Yin, M.-J., Yamamoto, Y. & Gaynor, R. B. *Nature* **396**, 77–80 (1998).
4. Israël, A. *Nature* **388**, 519–521 (1997).
5. Kopp, E. & Ghosh, S. *Science* **265**, 956–958 (1994).
6. Rainsford, K. D. *Aspirin and the Salicylates* (Butterworths, Boston, 1984).
7. Schwenger, P. et al. *Proc. Natl Acad. Sci. USA* **94**, 2869–2873 (1997).
8. Davis, R. J. *Trends Biochem. Sci.* **19**, 470–473 (1994).

Global change

A question of litter quality

Richard J. Norby and M. Francesca Cotrufo

A long-standing hypothesis has been laid to rest, one describing how terrestrial ecosystems will respond to rising levels of CO₂ in the atmosphere. The forum was a meeting in Capri*, where plant physiologists, ecologists and soil scientists met to re-examine the idea that changes in the chemistry of dead leaves (litter), grown in a high CO₂ environment, mean that its decomposition rate will decline. The consensus was that there is insufficient evidence to support this view. So it seems that what might have been a negative feedback on net plant productivity in the future, probably won't be.

The hypothesis was originally posed at a conference in 1982 (ref. 1). Plant tissue, including green leaves, was known to be lower in nitrogen concentration when grown in atmospheres enriched in CO₂, a response that has turned out to be quite consistent². If this signal is translated into the leaf litter entering the soil, then decomposing organisms would have food of lower quality, and the transfer of organically bound nitrogen to the pools of mineral nitrogen available for plant growth could slow. So the accelerat-

ed growth of plants in high CO₂ atmospheres would be self-limited (Fig. 1).

Although this has been an attractive idea — linking as it does a short-term, physiological response to a long-term response at a

larger, ecological scale — various studies have failed to support it. Most experiments, carried out in various ecosystems such as forests, agro-ecosystems, grasslands and a salt marsh, have reported little change in litter chemistry and no significant difference in decomposition rates under different CO₂ concentrations^{3–7} (A. Finzi, Duke Univ.; S. Hättenschwiler, Univ. Basel; B. Drake, Smithsonian Environmental Res. Center; C. Wood, Auburn Univ.). The discrepancy between the nitrogen concentrations in green leaves and leaf litter can be explained by the process of nutrient resorption during senescence, when leaf proteins and other nitrogenous compounds are hydrolysed and the products transported into perennial tissue before the leaves fall off the plant.

Rejecting the litter-quality hypothesis does not diminish the importance of other potential feedbacks on nitrogen availability (Fig. 1), or the need to characterize the processes that control decomposition in CO₂-enriched ecosystems. To evaluate the effect of rising CO₂ and associated climate variables on the sequestration of carbon, decomposition must be studied in relation to plant productivity and total litter input. For example, CO₂ enrichment of a Kansas prairie over an eight-year period resulted in a 6% increase in soil carbon content (J. Jastrow, Argonne Natl Lab.); this finding suggests that there is potential for increased carbon sequestration in the terrestrial biosphere. The carbon storage in this system was attributed to increased litter and increased physical protection of labile substrates from microorganisms that would otherwise decompose them. Detecting small increments in carbon storage in soil will require new techniques; the use of stable carbon isotopes holds particular promise in this respect.

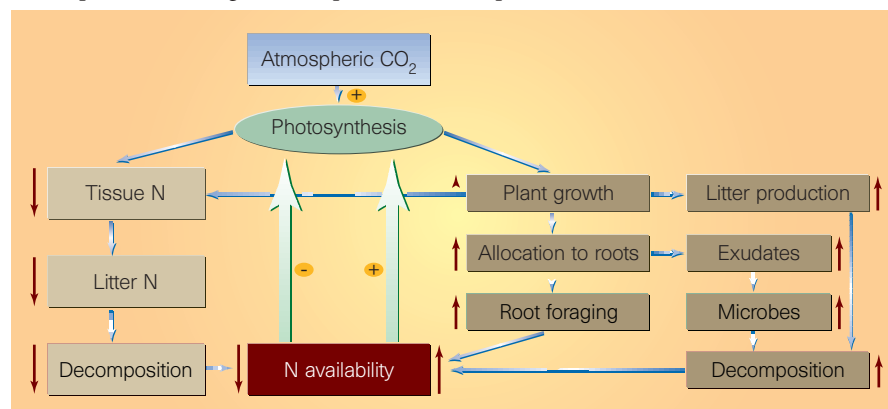


Figure 1 The primary effect of increasing atmospheric CO₂ in terrestrial ecosystems is to stimulate the rate of photosynthesis, but many other adjustments in plant and soil processes can result. The pathway on the left describes a negative feedback through changes in litter quality, decomposition and nitrogen availability. Experimental results do not provide convincing support for this pathway, and ecosystem models suggest that it is probably not important. On the right, other plant responses to high CO₂ that increase total litter production or root activity could provide a positive feedback on nitrogen availability. Decomposition and nitrogen availability will also be influenced by climatic factors, plant feedbacks on the soil environment, and nitrogen inputs and losses from the system. (Adapted from ref. 12.)

*Litter Quality and Decomposition under Elevated Atmospheric CO₂, Capri, Italy, 24–27 September 1998.

More sophisticated ecological approaches are also needed — the emphasis on changes in the leaf litter of individual plants has been too narrow. Shifts in plant community composition in response to high CO_2 can have dramatic effects on litter chemistry, and on carbon accumulation and nitrogen immobilization in soil (J. Dukes, Stanford Univ.). Changes induced by high CO_2 in the allocation of carbon to leaves, wood or roots will alter decomposition dynamics. Other environmental changes that will accompany the increasing concentration of CO_2 (warming and increased nitrogen deposition, for instance) will certainly influence decomposition rates, but their interactions with rising CO_2 are difficult to predict.

Another requirement is that experiments aiming to detect changes in the principal ecosystem properties that regulate nitrogen and carbon cycling must be guided by ecosystem models. The CENTURY model of decomposition⁸ indicates that litter chemistry does influence decomposition, but the changes that have been observed in experiments are simply too small to have a detectable effect on decomposition. Total litter input is much more important.

Other old ideas also need to be re-examined. For example, slower decomposition that immobilizes nitrogen in microbial biomass has been considered a negative feedback on productivity. But immobilization is a transient phenomenon and can be a 'good' thing — an indication of a more fertile habitat resulting from increased root litter in CO_2 -enriched grasslands⁹. And a surprising advance reported at the workshop is that slower initial decomposition rates can actually promote more complete long-term decomposition (B. Berg, Swedish Univ. Agric. Sci.); initial decomposition rate is promoted by high nitrogen concentration, but nitrogen retards long-term decomposition through inhibition of lignin-degrading enzymes and reactions that produce recalcitrant aromatic compounds¹⁰. Clearly, assessment of the carbon sequestration potential of soils in changing climates will require careful measurement of not just the litter inputs and decomposition, as well as other outputs from the ecosystem, but also analysis of the feedbacks between different processes and their temporal dynamics.

It is difficult to let favoured principles go. But the 15 years of work on the litter-quality hypothesis has led to a far richer understanding of how plants, decomposers and ecosystems will function as the atmosphere is progressively enriched with CO_2 and the climate changes. This understanding is especially needed as we struggle with the attempt to sequester the excess carbon we ourselves are emitting to the atmosphere¹¹. □

Richard J. Norby is in the Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6422, USA.

e-mail: rjn@ornl.gov

M. Francesca Cotrufo is in the Dipartimento di Scienze Ambientali, Seconda Università di Napoli, 81100 Caserta, Italy.

e-mail: fcotrufo@tin.it

1. Strain, B. R. & Bazzaz, F. A. in *CO₂ and Plants: The Responses of Plants to Rising Levels of Atmospheric Carbon Dioxide* (ed. Lemon, E. R.) 177–222 (Westview, Boulder, CO, 1983).
2. Cotrufo, M. F., Ineson, P. & Scott, A. *Glob. Change Biol.* **4**, 43–54 (1998).
3. O'Neill, E. G. & Norby, R. J. in *Carbon Dioxide and Terrestrial Ecosystems* (eds Koch, G. W. & Mooney, H. A.) 87–103 (Academic, San Diego, 1996).
4. Gahrooei, F. *Glob. Change Biol.* **4**, 667–677 (1998).

5. Torbert, H. A., Prior, S. A. & Rogers, H. H. *Soil Sci. Soc. Am. J.* **59**, 1321–1328 (1995).
6. Kemp, P. R., Waldecker, D. G., Owensby, C. E., Reynolds, J. F. & Virginia, R. A. *Pl. Soil* **165**, 115–127 (1994).
7. Franck, V. M., Hungate, B. A., Chapin, F. S. & Field, C. B. *Biogeochemistry* **36**, 223–237 (1997).
8. Parton, W. J. et al. *Glob. Biogeochem. Cycles* **7**, 785–809 (1993).
9. Thornley, J. H. M. (ed.) *Grassland Dynamics: An Ecosystem Simulation Model* (CAB International, Wallingford, Oxon, 1998).
10. Berg, B. et al. *Can. J. Bot.* **74**, 659–672 (1996).
11. IGBP Terrestrial Carbon Working Group *Science* **280**, 1393–1394 (1998).
12. Bazzaz, F. A. *Plants in Changing Environments: Linking Physiological, Population, and Community Ecology* (Cambridge Univ. Press, 1996).

Developmental biology

Sending all the right signals

Norbert Perrimon and Joseph B. Duffy

During growth and development, cells signal to — and regulate the fates of — each other. Cells can initiate many developmental fates in response to a single signal and, throughout the years, two mechanisms have been identified¹. In the first, receiving cells adopt different fates depending on the amount of signal or 'morphogen' that they receive. This implies that there is a concentration gradient of the signal, and that receiving cells can translate the quantitative amount of the signal they receive into different qualitative outcomes. The second mechanism is a 'relay mechanism', whereby receiving cells activate

secondary signals that are subsequently responsible for the diversification of cell fates.

Reporting in *Cell*, Wasserman and Freeman² provide a striking example of how a simple instructive signal initiates the development of an elaborate pattern during oogenesis. The oocyte (derived from the germ line) is surrounded by an epithelial layer of somatically derived follicle cells, which produce the chorion (or eggshell) in the final stages of oogenesis. Among the most prominent features of the chorion are two dorsal filaments, or appendages, used for respiration. In the fruitfly *Drosophila*

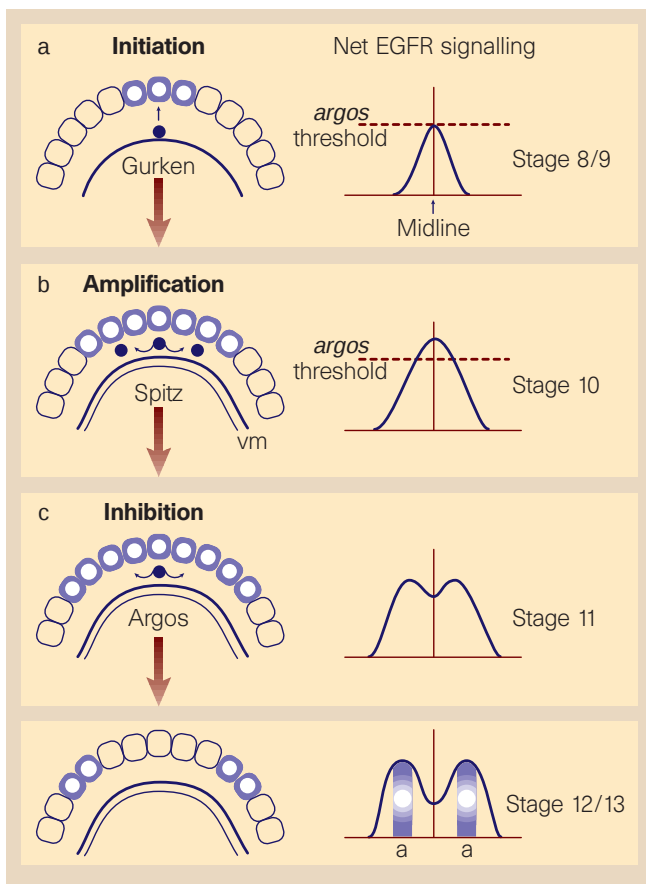


Figure 1 Cell-fate specification during dorsal patterning of the *Drosophila* egg. Wasserman and Freeman² have shown that patterning involves three stages of cell-fate specification. a, Activity of the epidermal growth factor receptor (EGFR) in the epithelial layer of follicle cells is initiated by a signal from the oocyte called Gurken. b, Activity of the EGFR is amplified through autocrine signalling — EGFR-mediated transcription of the *vein* gene, and of the *rho* gene, which increases the activity of another EGFR ligand, Spitz. c, EGFR also stimulates the transcription of *argos*, which inhibits the EGFR in the cells where it is most highly expressed. (Adapted from ref. 2.)

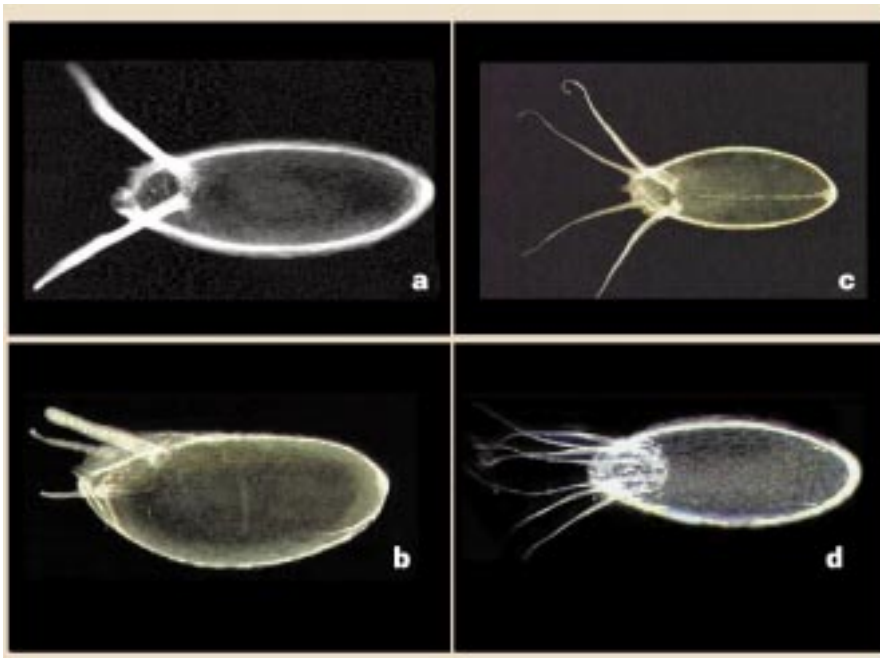


Figure 2 Diversity of dorsal appendages in *Drosophila* species. Chorions from: a, *D. melanogaster* (2 appendages); b, *D. phalerata phalerata* (3); c, *D. virilis* (4); d, *D. latifasciaeformis* (>4).

melanogaster, the fate of dorsal cells is specified by activation of the epithelial growth factor receptor (EGFR) in the follicle cells (see ref. 3 for review). Activation depends on an asymmetric signal encoded by Gurken (Grk), a transforming growth factor (TGF)- α -like molecule that originates from the oocyte. The Grk pathway initiates development of the respiratory appendages in the dorsal/anterior region of the chorion. But what is the nature of its contribution?

Wasserman and Freeman now document formation of the respiratory appendages by two distinct signalling mechanisms (Fig. 1). Initially, Grk activates EGFR in a paracrine fashion, through signalling between the oocyte and follicle cells. This establishes an initial profile of EGFR activity in a single, central domain within the dorsal anterior follicle cells. Paracrine signalling then activates a second phase of signalling during which the activity of EGFR is amplified in an autocrine fashion—the follicle cells signal to themselves.

The molecular details are as follows. In response to the initial phase of EGFR activity, transcription of the *rhomboid* (*rho*) and *vein* genes is induced in the follicle cells. The Vein protein, like Grk, encodes a putative ligand that can stimulate EGFR. The Rho protein, on the other hand, is not a ligand for EGFR, but it stimulates the activity of a third TGF- α -like ligand for EGFR, which is encoded by the *spitz* (*spi*) gene (see refs 4 and 5 for reviews). Thus, in response to the primary induction, a second phase of autocrine signalling involving two additional EGFR ligands leads to amplification of EGFR activity. But why? Wasserman and Freeman have an intriguing idea. During these final stages of

oogenesis, the follicle cells synthesize the vitelline membrane, which covers the oocyte and separates off the epithelial cells from the oocyte. This would block the activation of EGFR by the Grk signal from the oocyte. By maintaining the activation of EGFR in an autocrine fashion, however, this barrier is effectively circumvented.

How does this molecular activity result in the elaboration of two appendages? A third response to activity of the EGFR is transcription of the *argos* (*aos*) gene. In contrast to Grk, Vein and Spi, Aos encodes a secreted inhibitor of the EGFR. Wasserman and Freeman's findings indicate that Aos alters the initial profile of EGFR activity in the follicle cells. Aos acts within the peak of EGFR activity at the dorsal anterior, leading to repression along the centre of this profile. Effectively, repression splits the single peak into two (Fig. 1).

The authors confirmed this hypothesis by examining receptor-signalling activity in the following manner. Binding of ligand to the EGFR activates a cascade, involving the kinases Raf, mitogen-activated protein kinase kinase (MEK) and mitogen-activated protein kinase (MAPK), that has been conserved from invertebrates to vertebrates. MAPK is activated through phosphorylation by MEK at two sites. Antibodies that recognize this diphosphorylated form of MAPK can be used to examine the readout of signalling activity and, consistent with Wasserman and Freeman's theory, these antibodies indicated a single domain of activated MAPK in the dorsal anterior. This domain then resolves into two bilateral domains, marking the future position of the respiratory appendages.

Wasserman and Freeman propose that there are three stages of cell-fate specification: initiation of EGFR activity, its amplification, and subsequent repositioning. Furthermore, they document a hitherto unknown regulatory strategy—a transition from a paracrine signal to its subsequent autocrine amplification and inhibition. These findings have both evolutionary and ecological consequences. There are enormous diversities in the morphology, pattern and number of respiratory appendages in drosophilid chorions, ranging from no appendages to more than four (Fig. 2). These patterns seem to have evolved independently on numerous occasions⁶, presumably as a result of selective forces imposed by ecological constraints. With the molecular insight gained from *D. melanogaster*, we can now investigate the selective mechanisms that operate on a signal-transduction and pattern-forming pathway from molecular, evolutionary and ecological standpoints.

How has signalling been modified, and patterning differences achieved, over the course of evolution? One intriguing model is that diversity in the patterning of respiratory appendages has been achieved by increasing the complexity of EGFR regulation. Perhaps other positive and negative regulators of the EGFR remain to be identified. If so, such studies may have clinical implications. For example, misregulation of members of the EGFR/ErbB family of receptor tyrosine kinases has been implicated in more than 30% of human breast cancers, so by identifying new ways to inhibit these enzymes we may be able to develop cancer therapeutics.

Another model is that the diversity in appendage patterning has come about through cross-talk with other signalling pathways. Interestingly, the Decapentaplegic/TGF- β pathway is also involved in patterning of the appendages⁷ so, possibly, cross-talk between the EGFR and TGF- β pathway has been exploited to diversify patterning of the dorsal appendages. Regardless of how diversity in patterning is achieved, work in the fruitfly has once again proved to be ripe for the picking. □

Norbert Perrimon is in the Department of Genetics, Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts 02115, USA.

e-mail: perrimon@rascal.med.harvard.edu

Joseph B. Duffy is in the Department of Biology, Indiana University, Bloomington, Indiana 47405, USA.

e-mail: jduffy@bio.indiana.edu

1. Lawrence, P. A. & Struhl, G. *Cell* **85**, 951–961 (1996).
2. Wasserman, J. D. & Freeman, M. *Cell* **95**, 355–364 (1998).
3. Ray, R. P. & Schupbach, T. *Genes Dev.* **10**, 1711–1723 (1996).
4. Perkins, L. A. & Perrimon, N. *Cell* **89**, 13–16 (1998).
5. Schweitzer, R. & Shilo, B. Z. *Trends Genet.* **13**, 191–196 (1997).
6. Throckmorton, L. H. in *Studies in Genetics* 207–343 (Univ. Texas Press, 1962).
7. Twombly, V. et al. *Development* **122**, 1555–1565 (1996).

Jamming is not just cool any more

Andrea J. Liu and Sidney R. Nagel

All around us, things seem to be getting jammed. We travel on a highway and we are caught in traffic jams. At the wholefoods counter, grains and beans jam as they refuse to flow out of the bottom of the hopper into our bags. In factories, powdered raw materials clog the conduits that were designed to carry them smoothly. Our recourse in all these situations is to pound on our conduits, hoppers and dashboards until the jam miraculously disappears. We are usually so irritated that we have not really noticed that the jammed state, in all of these situations, has common properties. For example, the vibrations from the pounding actually do some good in reinitiating flow — except in the case of the traffic jam. Does the jammed solid then have different properties from the solids we normally encounter in the laboratory?

Writing in *Physical Review Letters*, Cates, Wittmer, Bouchaud and Claudin¹ contend that these jammed systems really belong to a new class of materials: 'fragile matter'. These systems resemble solids because the particles are driven into a jammed state by an externally applied stress. When jammed, the disordered system is caught in a small region of phase space with no possibility of escape.

Cates *et al.* propose that jammed systems are fundamentally different from ordinary solids in that, if the direction of the applied stress changes even by a small amount, then the jam will break up. A canonical example is a pile of sand, which appears solid: the upper surface slopes and sustains its shape despite the force of gravity, which one would expect to level the pile. But if one tilts or vibrates the pile, the grains shift and the solid melts. The authors argue that the unusual mechanical properties of fragile matter require a new theoretical description, which they first applied to a heap created by pouring sand onto the apex of a pile^{2,3}.

Traditionally, the forces within such a pile have been described using continuum elastoplastic theories. These are similar to models that describe ordinary solids⁴: every increment of stress in the material is related to a corresponding deformation, or strain⁵. The approach of Cates *et al.* is to start from a pile of completely non-deformable particles, for which strain is not an obviously useful variable. Their simple model of a chain of hard particles insists that the jammed system cannot be considered as an elastic body. Although it can support a large applied load in the same direction as the original jamming forces, the chain will fall apart if even an infinitesimal force is applied in a different direction. For an extended material such as a

sandpile, the material is fragile in the sense that a slight change in direction of the applied stress will change the entire structure of the force chains that give the pile its rigidity. Because there is no obvious relation connecting stress to strain throughout the pile, Cates *et al.* bypass the strain altogether and propose a relation between different components of the stress tensor^{2,3}. This continues to be a hotly debated assumption^{4,6–8}.

Cates *et al.* suggest that one way to reconcile the two approaches is to allow the particles to deform, so that the material can respond elastically to sufficiently small loads. One example of a system that is jammed and yet not fragile is foam. Shaving foam, for example, is jammed because the bubbles are tightly packed together under an isotropic stress, namely atmospheric pressure. If it were fragile, it would respond plastically to a shear stress, no matter how small. However, because bubbles deform, foam actually responds elastically as long as the stress is below a threshold value. Sand grains also deform slightly. Hence, for real systems, a continuum elastic description will always be useful. However, the new concept of fragile matter brings a valuable perspective from the opposite limit of completely non-deformable particles.

We would like to point out that the class of jammed materials may actually be broader than the authors suggest. They consider jamming only in systems with no attractive interactions (where the particle dynamics are constrained through an applied stress) and where the individual particles are large

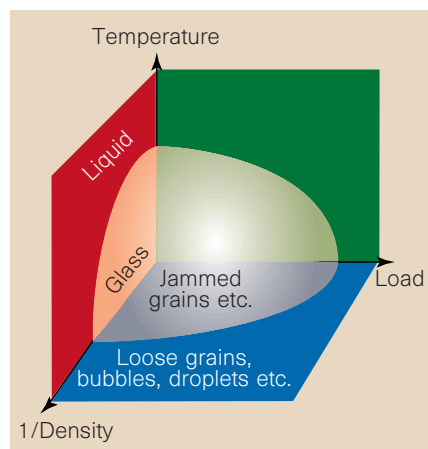


Figure 1 A possible phase diagram for jamming. The jammed region, near the origin, is enclosed by the depicted surface. The line in the temperature–load plane is speculative, and indicates how the yield stress might vary for jammed systems in which there is thermal motion.

so that there is no thermal motion. These two constraints may not be essential.

We know from studies of liquids and glasses that a system with attractive interactions often behaves in the same way as another that has only repulsive forces but is confined in a container (that constrains its density). In the case of jamming, the opposite situation may be possible: that is, one might be able to replace the constraints of an external pressure or stress with an attractive interaction between the particles. Thus, a supercooled liquid can be jammed into a glass simply by lowering the temperature, not by applying a stress. When a liquid is cooled below its freezing point, its viscosity increases rapidly. Eventually, it falls out of equilibrium into a disordered solid, or glass, where it only explores a small part of phase space, just as in the case of a jammed granular material or foam.

So might the concept of jamming and fragility include microscopic systems with attractive interactions, which unjam as one raises the temperature, as well as stressed macroscopic systems with repulsive interactions, which unjam as one applies an incompatible stress? We have sketched a speculative phase diagram for jamming (Fig. 1) that ties the different systems together. This phase diagram depends on temperature, load and density.

According to this picture, jamming can occur only when the density is high enough. One can then unjam the system either by raising temperature or by applying a stress. The phase diagram raises some interesting questions: for example, a glass may have a lower glass transition temperature under high shear stress. Likewise, a jammed granular material or foam may have a lower yield stress when random motions (that is, thermal fluctuations) are present. This would explain the beneficial role of banging on jammed conduits on the factory floor.

Whether jammed systems indeed share features that can be described by a phase diagram is an open question, but if our speculation has any merit it would bring together several different types of behaviour under one rubric. Are the dynamics of different systems approaching the jammed state also similar? If temperature and applied stress play similar roles in unjamming systems, is it possible that driven, macroscopic, athermal systems like granular materials and foams might be described in terms of an effective temperature? Is statistical mechanics useful at all in describing these systems? These and related questions will take years to resolve, but the picture of Cates *et al.* helps to point out some of the interesting conceptual problems that need to be addressed. □

Andrea J. Liu is in the Department of Chemistry and Biochemistry, University of California at Los Angeles, Los Angeles, California 90095-1569, USA. e-mail: liu@chem.ucla.edu

Sidney R. Nagel is at The James Franck Institute and the Department of Physics, University of Chicago, Chicago, Illinois 60637, USA.
e-mail: sid@rainbow.uchicago.edu

1. Cates, M. E., Wittmer, J. P., Bouchaud, J.-P. & Claudin, P. *Phys. Rev. Lett.* **81**, 1841–1844 (1998).
2. Bouchaud, J.-P., Cates, M. E. & Claudin, P. *J. Phys. I (Fr.)* **5**, 639–656 (1995).

3. Wittmer, J. P., Cates, M. E. & Claudin, P. *J. Phys. I (Fr.)* **7**, 39–80 (1997).
4. Savage, S. *Powders and Grains 97* (eds Behringer, R. P. & Jenkins, J. T.) 185–194 (Balkema, Rotterdam, 1997).
5. Cantelaube, F. & Goddard, J. D. *Powders and Grains 97* (eds Behringer, R. P. & Jenkins, J. T.) 231–234 (Balkema, Rotterdam, 1997).
6. Edwards, S. F. *Physica A* **249**, 226–231 (1998).
7. de Gennes, P. G. *Rev. Mod. Phys.* (submitted).
8. Narayan, O. & Nagel, S. R. *Physica A* (submitted).

Biodiversity

Plants on the web

David Read

What are the mechanisms that determine species composition in plant communities? In a world in which biodiversity decreases daily, the question is self-evidently important. One factor known to be central to the performance of individual plants, their interactions with each other, and even the species composition of the community itself, is the activity of mycorrhizal fungi, which are symbiotically associated with the roots of most land plants¹. Until now, however, most experiments designed to investigate these effects have employed either single² or undefined mixtures³ of fungal species, so their causes remain unclear.

A breakthrough comes with the study of van der Heijden *et al.* on page 69 of this issue⁴. Using field plots and mesocosms, the authors show first that the floristic diversity of two characteristically species-rich grass-

lands, one representative of calcareous habitats in Europe and the other of abandoned fields in North America, depends upon the presence of a species-rich assemblage of fungal symbionts in the soil. They then propose a mechanistic basis for the observed effect.

The importance of mycorrhizal fungi for nutrient acquisition by individual plants has been known for some time¹. But only recently have we become aware that because access to nutrients is increasingly restricted under competitive circumstances, mycorrhizal fungi can in part determine the outcome of interactions between plants, even to the extent of influencing the species composition of the community itself³. Emphasis upon the response of the plant has, however, left us with uncertainty about how the fungal partners bring about these changes. The unique feature of the work reported by van der Heijden *et al.* is that they have manipu-

lated the composition of the fungal population, and have observed the effects on communities of plant species representative of those normally found in these types of grassland.

When four different species of mycorrhizal fungi were added, either individually or in combination, to reconstructed calcareous grassland communities, it was observed that different fungal species promoted different plant hosts. The outcome was that alteration of the composition of the fungal taxa produced statistically significant changes in the structure and composition of the plant community. From such observations it can be hypothesized that, because of the additive beneficial effect of each single fungal species, plant biodiversity, and ecosystem productivity, will increase with numbers of fungal symbionts (Fig. 1).

Perhaps the most novel of all the experimental manipulations undertaken by the group involved a test of this hypothesis using the abandoned-field communities. Ecologists have long been interested in the factors determining plant species composition in these systems⁵. Here, instead of adding different fungal taxa to each of the standardized plant communities, 0 to 14 fungal species were added per treatment. As the number of fungal species per system was increased, the collective biomass of shoots and roots rose — as, most importantly, did the species diversity of the plant community.

What are the functional and mechanistic bases of these effects? The differential effects of specific plant–fungus combinations on growth of plant species may, as the authors suggest, provide one functional explanation, particularly as benefits to the plant would be expected to provide positive feedback to the fungal partner.

Differences in functional compatibility between plant and fungal species certainly appear to be at the heart of the observed effect. Similar differences have been reported previously but in laboratory studies of single plant–fungus combinations⁶. In these, the effects have been shown to be caused by fungus-specific differences in the ability to supply phosphorus to the plant.

However, the situation in nature is far more complex. Mycorrhizal fungi can enter the roots of most plant species. The outcome is not only that almost all plants in the community are simultaneously colonized by several species of mycorrhizal fungi, but also that they are interconnected by the external hyphal network of these fungi which forms a web extending from the root surface into the soil. Greater functional compatibility in nature must therefore arise when one fungus amongst a group occupying the root system enables one particular plant species to perform better than the others (Fig. 1). This could be facilitated by greater rates of phosphorus capture and transfer, by production

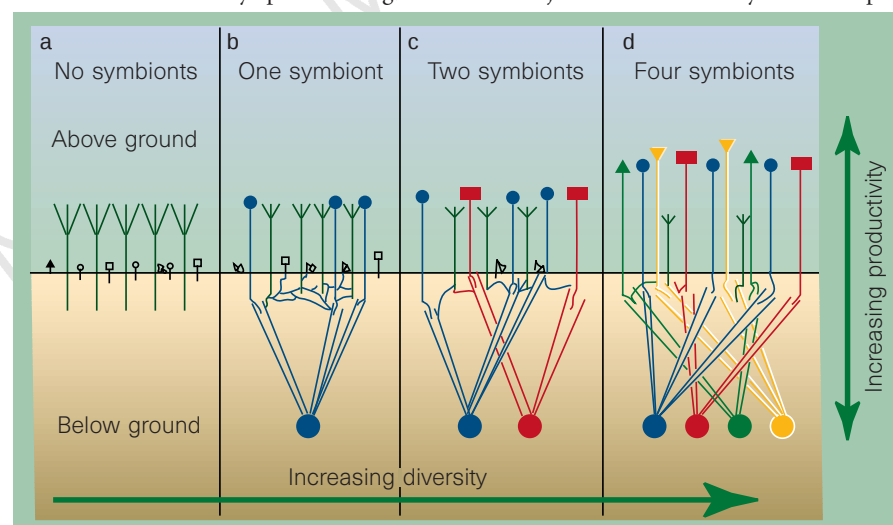


Figure 1 Possible basis for the effects of fungal species richness upon plant biodiversity and production observed by van der Heijden *et al.*⁴. Although most plant species are susceptible to colonization by mycorrhizal fungi, so becoming incorporated into a below-ground mycelial web, some of these fungi induce more positive responses in one plant species than in another. Because of these differences in functional compatibility, addition of new fungal species (here shown only up to four) leads to increases in survival and vigour of a progressively larger number of plant species. In the absence of symbionts (a), species that are relatively unresponsive to mycorrhizal fungi (grasses, for example) dominate. Addition of a fungal species, and then increase in their number (b–d) enables progressively more of the highly responsive herb species to develop, at the expense of the grasses. Positive feedback to the increasingly diverse fungal population may lead to greater vigour of the fungal web, so contributing to more effective exploitation of soil resources and greater overall productivity.

of a larger 'sink' for carbon or by a combination of both effects.

Irrespective of its physiological basis, it is clear that differences in functional compatibility can explain both the stimulatory effect on individual plant species and the influence of fungal species diversity upon the overall structure of the plant community. In the experimental system of van der Heijden *et al.*⁴, plant diversity reached a maximum when eight fungal species were added — a number that must in part reflect the physical and biological constraints of the authors' reconstructed communities. It would take a brave person to propose a likely optimal number under field conditions, and a rash one to suggest, at this stage of our understanding, that any symbiont might be redundant.

Although debate continues over the contribution of diversity to ecosystem function, empirical studies provide support for the view that floristically rich systems are more productive⁷, show greater stability under stress⁸ and are more likely to provide alleviation of global problems posed by atmospheric CO₂ enrichment⁹. A recognition of these properties, coupled with an increasing awareness that the diversity of terrestrial vegetation systems is everywhere under threat, has forced biologists to consider the mechanisms that determine species composition in plant communities.

The study of van der Heijden *et al.* high-

lights the essentially interactive nature of those mechanisms. It also shows that conservation of the fungal gene pool is likely to be a prerequisite for maintenance of floristic diversity in grasslands, as well as in other ecosystems such as boreal forests¹⁰, where the fungal web is known to influence allocation of resources between plant species. We are already aware of many of the sensitivities of the mycorrhizal community to perturbations, particularly those associated with cultivation and nutrient enrichment¹¹. So the new findings may be of direct practical importance for those managers who seek to optimize floristic diversity, be it for environmental, conservation or amenity purposes. □

David Read is in the Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK.

e-mail: D.J.Read@sheffield.ac.uk

1. Smith, S. E. & Read, D. J. *Mycorrhizal Symbiosis* 2nd edn (Academic, San Diego, 1997).
2. Fitter, A. H. *New Phytol.* **79**, 119–125 (1977).
3. Grime, J. P., Mackey, J. M. L., Hillier, S. H. & Read, D. J. *Nature* **328**, 420–422 (1987).
4. van der Heijden, M. G. A. *et al.* *Nature* **396**, 69–72 (1998).
5. Whittaker, R. H. *Communities & Ecosystems* 2nd edn (Macmillan, New York, 1975).
6. Ravnkov, S. & Jakobsen, I. *New Phytol.* **129**, 611–618 (1995).
7. Tilman, D., Wedin, D. & Knops, J. *Nature* **379**, 718–720 (1996).
8. Tilman, D. & Downing, J. A. *Nature* **367**, 363–365 (1994).
9. Naeem, S., Thompson, L. J., Lawler, S. P., Lawton, J. H. & Woodfin, R. M. *Nature* **368**, 734–737 (1994).
10. Simard, S. W. *et al.* *Nature* **388**, 579–582 (1997).
11. Helgason, T., Daniell, T. J., Husband, R., Fitter, A. H. & Young, J. P. Y. *Nature* **394**, 431 (1998).

Telomeres

Bumps on the road to immortality

Robert A. Weinberg

In contrast to normal cells, which double only a limited number of times in culture, almost all malignant tumour cells can double indefinitely. Telomeres — specialized structures at the end of chromosomes — seem to be central molecular determinants of this immortality trait, and their role is addressed by Kiyono *et al.*¹ on page 84 of this issue.

Telomeric DNA is essential to preserve the ends of chromosomes and, hence, the integrity of the genome. According to one model², in normal cells telomeres shorten with each generational doubling. Ultimately they erode to a size that causes cells to senesce or, if the cells manage to circumvent this initial blockade, to enter into crisis (when karyotypic instability accompanies, and probably causes, widespread death in cell populations). But expression of telomerase, an enzyme that is repressed in most post-embryonic human cells, allows cells to repair their telomeres. In this way cancer cells can subvert the generational clock, avoiding senescence or crisis, and achieve unlimited proliferation.

This model received strong support from a landmark paper³ published this January, in

which the ectopic expression of *hTERT* — the catalytic subunit of the telomerase holoenzyme — enabled human cells that were destined to senesce to multiply indefinitely^{3,4}. Thus, ectopic expression of *hTERT* seems to have allowed these cells — retinal pigmented epithelial (neuorectodermal) cells and fibroblasts — to avoid both senescence and the subsequent crisis.

This result was a Holy Grail for many cell biologists, who have wrestled unsuccessfully with the difficulties of propagating differentiated cells from various tissues in culture. Almost invariably such cells stop growing — victims of senescence. To try and immortalize primary cells, some workers have introduced viral oncogenes such as the SV40 large T antigen oncogene into them. Indeed, SV40 large T is known to allow cells to circumvent senescence⁵. However, although this trick may extend life span, the resulting cell cultures often represent a pyrrhic victory because the immortalizing agent also perturbs their differentiation programmes. In contrast, ectopic expression of telomerase promises to confer replicative immortality on



100 YEARS AGO

The introduction of electricity into our houses has added materially to the comfort and luxury of home. If we were living in the days of ancient Greece, the presiding domestic deity would have been *Electra*. The old bellhanger has been rung out by the new goddess. *Electra* ... attracts the attention of our domestics, not by a gamut of ill-toned and irregularly-excited bells, but by neat indicators and one uniform sound. The timid visitor fears no more that he has expressed rage or impatience by his inexperience of the mechanical pull required at the front door. The domestic telephone is coming in as an adjunct to the bell. Its use saves two journeys. The bell attracts attention, the telephone transmits the order ... Heating appliances are becoming very general ... Radiators assist the coal fire by maintaining the temperature of a room uniform throughout its length and breadth. Ovens are heated, water is boiled, flat-irons become and are maintained at a useful temperature, breakfast dishes and tea-cakes are kept hot, even curling-tongs have imparted to them the requisite temperature to perform their peculiar function.

From *Nature* 3 November 1898.

50 YEARS AGO

The award of the Nobel Prize for Medicine for 1948 has been made to Dr. Paul Müller for his discovery of the effects as an insecticide of D.D.T. Drs. P. Läger, P. Müller and H. Martin were the leaders of an intensive research for insecticidal chemicals in the Basle laboratories of J. R. Geigy, S.A., which extended more than twenty years. The researches were directed originally towards the discovery of moth-proofing agents ... It seems to have been in the course of field trials that its remarkable effectiveness against the Colorado beetle was noticed. It was soon found to be equally toxic to the louse and the mosquito. The material was brought to the notice of medical entomologists ... at a critical moment in the War when the supplies of pyrethrum were rapidly falling short of the demand ... It proved of enormous value in combating typhus and malaria during the War, and now it is being employed with success in campaigns for the complete eradication of malaria from island areas.

From *Nature* 6 November 1948.

cells without affecting other regulatory systems such as those that control differentiation.

But the work of Kiyono *et al.*¹ now indicates that it may not be so easy to avoid senescence, at least in certain cell types. Their work with two human epithelial cell types — keratinocytes and mammary epithelial cells — indicates that immortalization of these cells requires expression of the *hTERT* gene and inactivation of the retinoblastoma (Rb)/p16^{INK4a} tumour-suppressor pathway. This pathway halts cells in the G1 phase of the cell cycle in response to a range of physiological signals known to inhibit growth (Fig. 1).

These results highlight the complexity of the senescence phenotype. The very name is misleading, because senescence is provoked by many stimuli that have nothing to do with cell ageing. For example, activation of oncogenes can induce a cell to senesce rapidly⁶. Hence, telomere shortening cannot be the only stimulus to provoke senescence. This conclusion is made obvious by the observation that when primary cells are placed into culture they often senesce within a small number of generations — long before their telomeres have eroded significantly.

Such observations might indicate that the senescence phenotype signals the difficulties that cells experience when they are explanted from living tissue into the artificial environment of the culture dish. Physiological stress may, for example, activate the p16^{INK4a} cyclin-dependent kinase inhibitor, which in turn shuts down growth by blocking phosphorylation of Rb. Hence, senescence provoked by dif-

ficulties in adapting to a culture medium may be avoided only by neutralizing the p16^{INK4a}/Rb braking system. Kiyono *et al.*¹ exploit human papillomavirus oncoproteins to do so, but the SV40 large T oncoprotein used by others seems to achieve the same result⁵.

All this converges on the idea that the epithelial cells studied by Kiyono *et al.* — and, perhaps, all epithelial cells — may need to overcome the senescence induced by explantation into culture before they confront the barrier to unlimited proliferation of telomere shortening. Such a sequence of events might explain why inactivation of the Rb pathway and ectopic expression of telomerase are both required for immortalization of epithelial cells. For those intent on using *hTERT* to immortalize primary epithelial cells and study their differentiation phenotypes, this model, if extended and validated, will have unfortunate consequences because inactivation of the Rb pathway often affects differentiation programmes⁷.

This work may also affect our view of how tumours progress in the 90% of human cancers that derive from epithelia. Cell immortalization may be a two-step process at the least. The first step, inactivation of the p16^{INK4a}/Rb pathway, may occur relatively early, enabling cells to avoid the senescence provoked by, for example, oncogene activation⁶. Only later in tumour progression, when clones of premalignant cells have begun to exhaust their endowment of telomeres, will activation of telomerase through derepression of *hTERT* become advantageous for further proliferation.

Moreover, the apparent separation between oncogenes, tumour-suppressor genes and *hTERT* may be an illusion. A recent study⁸ indicates that the *myc* oncogene upregulates telomerase expression, and the work of Kiyono *et al.*¹ and earlier work of others⁹ shows that the E6 oncoprotein of the human papillomavirus can induce expression of *hTERT*. Accordingly, the circuitry that governs the cell-cycle clock may be tightly linked to that governing the telomere-based generational clock. Thus, oncogene activation may yield both short-term advantages in deregulating advance of the cell cycle and long-term advantages by stabilizing telomeres. □

Robert A. Weinberg is at the Whitehead Institute for Biomedical Research, and the Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA.
e-mail: weinberg@wi.mit.edu

1. Kiyono, T. *et al.* *Nature* **396**, 84–88 (1998).
2. Sedivy, J. M. *Proc. Natl Acad. Sci. USA* **95**, 9078–9081 (1998).
3. Bodnar, A. G. M. *et al.* *Science* **279**, 349–352 (1998).
4. Vaziri, H. & Benchimol, S. *Curr. Biol.* **8**, 279–282 (1998).
5. Wright, W. E. & Shay, J. W. *Exp. Gerontol.* **27**, 383–389 (1992).
6. Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. *Cell* **88**, 593–603 (1997).
7. Jacks, T. & Weinberg, R. A. *Science* **280**, 1035–1036 (1998).
8. Wang, J., Xie, L. Y., Allan, S., Beach, D. & Hannon, G. J. *Genes Dev.* **12**, 1769–1774 (1998).
9. Klingelutz, A. J., Foster, S. A. & McDougall, J. K. *Nature* **380**, 79–82 (1996).

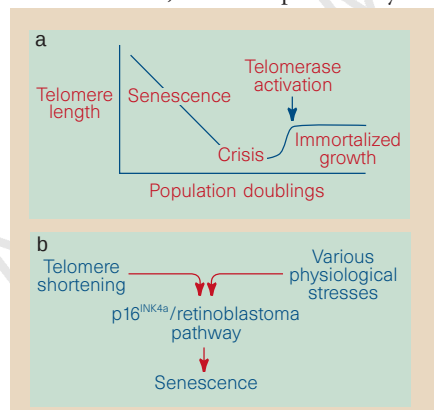


Figure 1 Models for the role of telomeres in cell immortalization. a, Standard model. With each cell division, the telomeres at the end of the chromosomes become shortened. In normal cells this leads to senescence and eventually to crisis. But activation of telomerase, in cancer cells for example, allows the telomeres to be regenerated indefinitely, immortalizing these cells. b, Revised model, according to the results of Kiyono *et al.*¹. They show that activation of telomerase alone is not enough to immortalize certain epithelial cells, and that inactivation of the p16^{INK4a}/retinoblastoma pathway is also needed. Because this pathway is activated by many other physiological stresses, immortalizing cells may not be as simple as was previously thought.

Daedalus

Smokeless microwaves

A flame keeps burning by thermal feedback. The heat it gives out vaporizes incoming fuel, and raises it to combustion temperature. Sadly, the process is not particularly efficient. Vaporized fuel is never completely burnt in its passage through the flame. A little unburnt fuel, and a variety of partly burnt combustion products, always escape into the cooling flue gases. The result is soot, smoke, and frequently the need for a costly catalytic converter downstream, to oxidize the pollutants before they get away.

So Daedalus is planning an alternative form of energetic feedback. His idea is to beam intense microwaves into the burner or furnace. The hot, ionized combustion gases should be conducting enough to absorb the energy, which will make them hotter still. The beam should be concentrated just above or beyond the flames, to heat and oxidize the partially burnt gases before they can cool down. Visible smoke, in the form of carbonaceous particles, will be particularly well suppressed. Carbon is such a good conductor that the microwaves will keep it glowing until it has burnt completely. But microwaves seem to encourage chemical reactions not merely by heating, but also by direct molecular agitation. Nasties such as carbon monoxide and nitric oxide absorb them strongly, and should be rapidly reacted away.

The obvious application is to big power-stations. They could well afford to boost their combustion efficiency by feeding a bit of power back into their furnaces. Old, polluting coal-fired stations, and ecologically virtuous ones heroically trying to burn rubbish, would benefit the most. Indeed, even the Greens might applaud a guaranteed pollution-free, microwave-augmented incinerator.

The internal-combustion engine is a more daring challenge. DREADCO engineers are now plumbing microwave waveguides into the cylinders of a test diesel engine. A dynamo on the engine will power a klystron. At each moment of ignition, it will beam an intense microwave pulse into the firing cylinder, which will act as a frequency-swept resonant cavity. Smoke and partial-combustion products should vanish. They will be burned, not in a wasteful exhaust catalyser, but in the engine itself, thus raising its efficiency. Even the microwave energy beamed into the cylinders will boost the engine's power, and be partly returned to the dynamo generating it.

David Jones

Vincian Velcro

Leonardo demonstrated his innovatory drawing techniques in his sketch of a human fetus. The work was partly inspired by his studies of botany and his dissection of a cow — and it hints at his thoughts on the nature of life itself.

Martin Kemp

All the new graphic modes forged during the Renaissance were realized at the highest level in the drawings of Leonardo da Vinci. Indeed, he invented key techniques, including the exploded view and solid section for the disclosing of natural structures.

His famous drawing of a fetus in the womb is a miracle of intense presentation. The womb, split open like a burst seed-case, reveals the coiled fetus, shaped into compelling roundness by the rhythmic curves of his pen.

The botanical analogy is typical of Leonardo's search for microcosmic parallels between all created forms, and is underscored by the ancillary sketches in the lower centre of the sheet, where successive membranes of the womb are peeled away layer by layer to disclose its tender secrets.

The womb is clearly not that of a human mother. The cotyledonous placenta, composed from separate pads attached across the surface of the womb, has been adapted from the beautiful drawing he made of his dissection of the gravid uterus of a cow. The perfection of design that he found in that dissection could not, in his mind, be other than present in the human womb, complementing the innate sphericity of the chamber in which the new life germinates.

Alternative configurations for the interdigitations of the 'male and female parts' of the cotyledons are envisaged in two brilliantly innovatory drawings to the upper right. The peeling back of one of the pads, separating reluctantly like a Velcro strip, is a graphic device of extraordinary eloquence.

On first glance, the two sketchy diagrams on the sheet, to the centre right and in the bottom right corner, appear to be unrelated — apparently typical of the magpie restlessness of Leonardo's thought. However, the former, which demonstrates how a sphere may be stabilized on an inclined plane by an eccentric weight, probably arose during his meditations on the orientation of the fetus, with its heavy head in its bath of fluid. And the latter, which demonstrates "why a picture seen with one eye will not demonstrate such relief as the object seen with two eyes", serves as a characteristic reminder that even his skills cannot equal the effect of the actual form seen with binocular sight.

But we are still missing the core of Leonardo's vision. In the note below the



Leonardo da Vinci's "The infant and womb", in the Royal Library at Windsor Castle.

main drawing, he speculates on a matter that we would not deem to be susceptible to 'anatomical' investigation: "The same soul governs these two bodies, and the desires and fears and sorrows are common to this creature as to all the other animated parts, and from this it arises that something desired by the mother is often found engraved on the limbs of the infant... and a sudden terror destroys the mother and the child."

Viewed in this light, the drawing is a study of what he calls the "great mystery" of

the fetus in the waters of the womb, an unawakened life whose thoughts and body are being irresistibly shaped by a oneness of two beings that can never be equalled. There was no such thing as 'descriptive anatomy' for Leonardo, unless we extend description to embrace the profoundest instincts about human life in the greater scheme of things. □

Martin Kemp is in the Department of the History of Art, University of Oxford, 35 Beaumont Street, Oxford OX1 2PG, UK.

e-mail: martin.kemp@trinity.oxford.ac.uk

Jefferson fathered slave's last child

There is a long-standing historical controversy over the question of US President Thomas Jefferson's paternity of the children of Sally Hemings, one of his slaves¹⁻⁴. To throw some scientific light on the dispute, we have compared Y-chromosomal DNA haplotypes from male-line descendants of Field Jefferson, a paternal uncle of Thomas Jefferson, with those of male-line descendants of Thomas Woodson, Sally Hemings' putative first son, and of Eston Hemings Jefferson, her last son. The molecular findings fail to support the belief that Thomas Jefferson was Thomas Woodson's father, but provide evidence that he was the biological father of Eston Hemings Jefferson.

In 1802, President Thomas Jefferson was accused of having fathered a child, Tom, by Sally Hemings⁵. Tom was said to have been born in 1790, soon after Jefferson and Sally Hemings returned from France where he had been minister. Present-day members of the African-American Woodson family believe that Thomas Jefferson was the father of Thomas Woodson, whose name comes from his later owner⁶. No known documents support this view.

Sally Hemings had at least four more children. Her last son, Eston (born in 1808), is said to have borne a striking resemblance to Thomas Jefferson, and entered white society in Madison, Wisconsin, as Eston Hemings Jefferson. Although Eston's descendants believe that Thomas Jefferson was Eston's father, most Jefferson scholars give more credence to the oral tradition of the descendants of Martha Jefferson Randolph, the president's daughter. They believe that Sally Hemings' later children, including Eston, were fathered by either Samuel or Peter Carr, sons of Jefferson's sister, which would explain their resemblance to the president.

Because most of the Y chromosome is passed unchanged from father to son, apart from occasional mutations, DNA analysis of the Y chromosome can reveal whether or not individuals are likely to be male-line relatives. We therefore analysed DNA from the Y chromosomes of: five male-line descendants of two sons of the president's paternal uncle, Field Jefferson; five male-line descendants of two sons of Thomas Woodson; one male-line descendant of Eston Hemings Jefferson; and three male-line descendants of three sons of John Carr, grandfather of Samuel and Peter Carr (Fig. 1a). No Y-chromosome data were available from male-line descendants of President Thomas Jefferson because he had no surviving sons.

Seven bi-allelic markers (refs 7-12), eleven microsatellites (ref. 13) and the minisatellite MSY1 (ref. 14) were analysed (Fig. 1b). Four of the five descendants of Field

Jefferson shared the same haplotype at all loci, and the fifth differed by only a single unit at one microsatellite locus, probably a mutation. This haplotype is rare in the population, where the average frequency of a microsatellite haplotype is about 1.5 per cent. Indeed, it has never been observed outside the Jefferson family, and it has not been found in 670 European men (more than 1,200 worldwide) typed with the microsatellites or 308 European men (690 worldwide) typed with MSY1.

Four of the five male-line descendants of Thomas Woodson shared a haplotype (with one MSY1 variant) that was not similar to the Y chromosome of Field Jefferson but was characteristic of Europeans. The fifth Woodson descendant had an entirely different haplotype, most often seen in sub-Saharan Africans, which indicates illegitimacy in the line after individual W42. In contrast, the descendant of Eston Hemings Jefferson did

have the Field Jefferson haplotype. The haplotypes of two of the descendants of John Carr were identical; the third differed by one step at one microsatellite locus and by one step in the MSY1 code. The Carr haplotypes differed markedly from those of the descendants of Field Jefferson.

The simplest and most probable explanations for our molecular findings are that Thomas Jefferson, rather than one of the Carr brothers, was the father of Eston Hemings Jefferson, and that Thomas Woodson was not Thomas Jefferson's son. The frequency of the Jefferson haplotype is less than 0.1 per cent, a result that is at least 100 times more likely if the president was the father of Eston Hemings Jefferson than if someone unrelated was the father.

We cannot completely rule out other explanations of our findings based on illegitimacy in various lines of descent. For example, a male-line descendant of Field

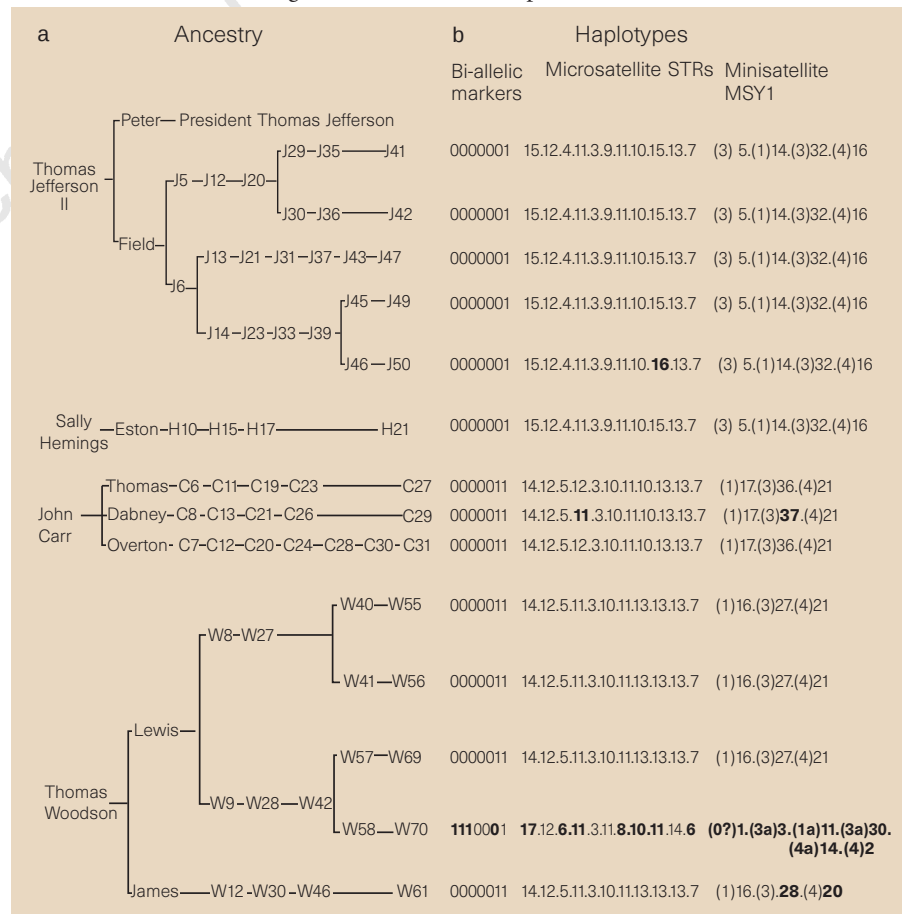


Figure 1 Male-line ancestry and haplotypes of participants. **a**, Ancestry. Numbers correspond to reference numbers and names in more detailed genealogical charts for each family. **b**, Haplotypes. Entries in bold highlight deviations from the usual patterns for the group of descendants. **Bi-allelic markers**. Order of loci: YAP-SRYm8299-sY81-LLY22g-Tat-92R7-SRYm1532. 0, ancestral state; 1, derived state. **Microsatellite short tandem repeats (STRs)**. Order of loci: 19-388-389A-389B-389C-389D-390-391-392-393-dxys156y. The number of repeats at each locus is shown. **Minisatellite MSY1**. Each number in brackets represents the sequence type of the repeat unit; the number after it is the number of units with this sequence type. For example, J41 has 5 units of sequence type 3, 14 units of sequence type 1, 32 units of sequence type 3, and 16 units of sequence type 4.

Jefferson could possibly have illegitimately fathered an ancestor of the presumed male-line descendant of Eston. But in the absence of historical evidence to support such possibilities, we consider them to be unlikely.

Eugene A. Foster*, **M. A. Jobling†**,
P. G. Taylor†, **P. Donnelly‡**, **P. de Knijff§**,
Rene Mieremet§, **T. Zerjal¶**, **C. Tyler-Smith¶**

*6 Gildersleeve Wood, Charlottesville,
Virginia 22903, USA

e-mail: eafoster@aol.com

†Department of Genetics,
University of Leicester,
Adrian Building, University Road,
Leicester LE1 7RH, UK

‡Department of Statistics,
University of Oxford,
South Parks Road, Oxford OX1 3TG, UK

§MGC Department of Human Genetics,
Leiden University, PO Box 9503,
2300 RA Leiden, The Netherlands

¶Department of Biochemistry,
University of Oxford,
South Parks Road, Oxford OX1 3QU, UK

1. Peterson, M. D. *The Jefferson Image in the American Mind* 181–187 (Oxford Univ. Press, New York, 1960).
2. Malone, D. *Jefferson the President: First Term, 1801–1805* Appendix II, 494–498 (Little Brown, Boston, MA, 1970).
3. Brodie, F. M. *Thomas Jefferson: An Intimate History* (Norton, New York, 1974).
4. Ellis, J. J. *American Sphinx: The Character of Thomas Jefferson* (Knopf, New York, 1997).
5. Callender, J. T. *Richmond Recorder* 1 September 1802. [Cited in: Gordon-Reed, A. *Thomas Jefferson and Sally Hemings: An American Controversy* (Univ. Press of Virginia, Charlottesville, 1997)].
6. Woodson, M. S. *The Woodson Source Book* 2nd edn (Washington, 1984).
7. Hammer, M. F. *Mol. Biol. Evol.* **11**, 749–761 (1994).
8. Whitfield, L. S., Sulston, J. E. & Goodfellow, P. N. *Nature* **378**, 379–380 (1995).
9. Seielstad, M. T. et al. *Hum. Mol. Genet.* **3**, 2159–2161 (1994).
10. Zerjal, T. et al. *Am. J. Hum. Genet.* **60**, 1174–1183 (1997).
11. Mathias, N., Bayes, M. & Tyler-Smith, C. *Hum. Mol. Genet.* **3**, 115–123 (1994).
12. Kwok, C. et al. *J. Med. Genet.* **33**, 465–468 (1996).
13. Kayser, M. et al. *Int. J. Legal Med.* **110**, 125–133 (1997).
14. Jobling, M. A., Bouzekri, N. & Taylor, P. G. *Hum. Mol. Genet.* **7**, 643–653 (1998).

fall into two classes, according to whether the vacuoles in medullary keratin scatter light coherently (with a relationship between the phases of waves scattered by different surfaces) or incoherently. In the Rayleigh and Mie scattering models^{4,6}, the colours are produced by incoherent scattering of smaller visible wavelengths by an array of spatially independent air vacuoles in the medullary keratin. Alternatively, in the constructive interference model, the colours are produced by interactions between light waves scattered coherently by the keratin–air matrix^{1–3,5}. Out-of-phase waves will destructively interfere and cancel out, whereas in-phase waves will constructively reinforce one another and be reflected coherently. The phase relationships of scattered waves are determined by the size and spatial distribution of the scatterers^{1,2,7}. Periodic spatial relationships between scatterers will produce consistent path-length differences among scattered waves and reinforce a limited set of wavelengths.

To investigate how the structural colour of feather barbs is produced, we apply an electromagnetic theory of coherent light scattering⁸ that was developed to explain the transparency of the human cornea. In a

quasi-random array of scatterers, coherent scattering is predicted only for light waves with a wavelength of twice that of the largest components of the Fourier transform of the spatial variation in refractive index⁷. We performed a discrete 2D Fourier analysis of digitized transmission electron micrographs of cross-sections of the medullary keratin from the blue feather barbs of *C. maynana* to examine whether it has periodic spatial variation in refractive index. The spongy medullary layer of *C. maynana* barbs consists of a matrix of keratin bars with air vacuoles of strikingly uniform diameter and spacing². The matrix has no planes of symmetry and is randomly orientated to the surface of the feather. The 2D Fourier power spectrum of the medullary keratin matrix shows a nearly circular ring around the origin 0.0059 nm^{-1} in diameter (Fig. 1a). This ring indicates that the tissue has a highly ordered, nanoscale spatial variation in refractive index that is nearly uniform in all directions.

Reflection spectra of the blue feather barbs display a peak between 500 and 520 nanometres (Fig. 1b). We used the 2D Fourier power spectrum to predict the reflectance spectrum of these barbs. The

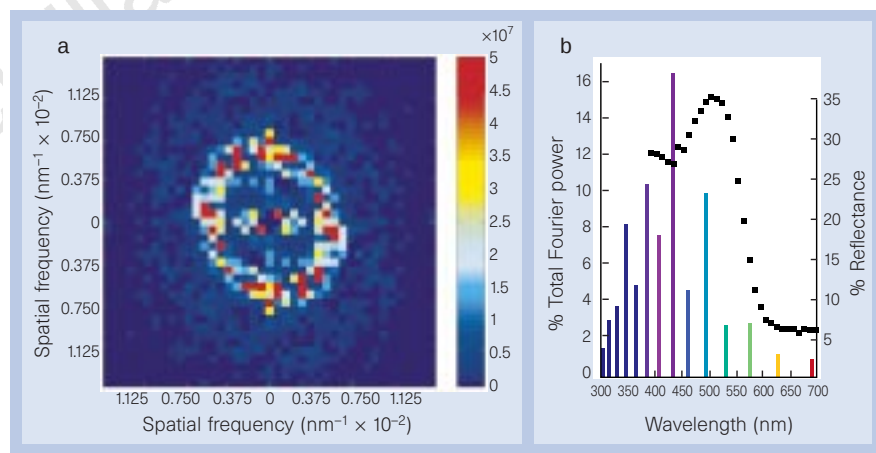


Figure 1 Two-dimensional Fourier power spectrum and observed and predicted reflectance spectra of a blue feather barb of *Cotinga maynana*. **a**, 2D Fourier power spectrum of the spongy medullary keratin matrix of a blue feather barb. **b**, Observed reflectance spectrum of a blue feather barb (black squares, right axis) and predicted reflectance spectrum (bars, left axis) based on the 2D Fourier power spectrum (**a**). The discrete Fourier transform describes how a signal or image is composed of different periodic component waves⁹. A 2D Fourier analysis was conducted using the 2D Fast Fourier Transform algorithm in MATLAB[®] on a digitized 500-pixel² image of a transmission electron micrograph (magnification, $\times 29,000$)². The image was processed in MATLAB using variance rescaling with block processing and median filtering¹⁰. The 2D Fourier power spectrum (**a**) resolves the spatial variation in optical density in the tissue into its periodic components in any direction in the plane from all points. The size and direction of any component spatial frequency is given by the length and direction of a vector from the origin to that point. The squared amplitude of the component is given by the colour (scale bar on the right). The ring indicates a highly ordered, nanoscale fluctuation in optical density moving in any direction in the plane of the tissue at a spatial frequency of about 0.0059 nm^{-1} . Reflectance of blue feather barbs (**b**, squares, right axis) was measured using a Zeiss microspectrophotometer at 30 wavelengths between 400 and 700 nm¹². The predicted reflectance spectrum (**b**, bars, left axis) expresses the percentage of the total power distributed among the components of the 2D Fourier power spectrum over all directions. The predicted reflection spectrum was calculated using the radial average of one quadrant of the 2D Fourier power spectrum with half-pixel width intervals, normalizing the volume under the rotated radial average function to one, calculating the volume under radial sections of the radial average Fourier power function, multiplying these volumes by twice the mean refractive index (1.283), estimated from the micrographs using the refractive indices of keratin (1.54) and air (1.00)⁸, and plotting these values as wavelengths in nanometres.

Coherent light scattering by blue feather barbs

The structural colours of avian feather barbs are created by the scattering of light from the spongy matrix of keratin and air in the medullary layer of the barbs^{1–5}. However, the precise physical mechanism for the production of these colours is still controversial^{1,3,4,6}. Here we use a two-dimensional (2D) Fourier analysis of the spatial variation in refractive index of the blue feather barbs of the plum-throated cotinga (*Cotinga maynana*, Cotingidae) to show that the colour is produced by constructive interference between light waves scattered coherently by the nanostructured keratin–air matrix of the barbs.

Mechanisms proposed to explain the production of structural colours of feather barbs

predicted reflectance spectrum shows a peak at 435 nanometres with descending slopes towards both ends of the visible spectrum (Fig. 1b), which agrees well with the observed reflectance spectrum in hue (position of peak) and chroma (shape of peak). Most of the total power in the Fourier power spectrum is concentrated at spatial frequencies that will produce coherent scattering of light waves that are close to the observed blue hue. The correspondence between the peaks of the observed and predicted reflection spectra (about 75 nanometres) is striking as there are no reasons, apart from structural colour production, to expect these spatial frequencies to predominate.

In the Mie and Rayleigh scattering models, the medullary keratin scatters incoherently (which means the scatterers must be separated by distances larger than the wavelength of visible light)⁸. However, the ring-like Fourier power spectrum demonstrates that the scatterers in medullary keratin are spatially periodic on a scale smaller than visible wavelengths, so scattering is coherent. Furthermore, the air vacuoles in the medullary keratin of *C. maynana* and other structurally coloured feather barbs^{3,5} are too close together to scatter visible light incoherently. Finally, the Rayleigh model predicts that the reflectance spectrum should increase continuously into the ultraviolet¹, but the distinct peak in the reflection spectrum shows this prediction to be false (Fig. 1b). Rayleigh scattering has been the traditional explanation of this phenomenon for nearly a century⁴, and remains widely cited in ornithological texts and reference works⁹, but our results indicate that the Rayleigh model and other incoherent scattering models are not suitable explanations.

Our analysis of the spatial variation in the refractive index of medullary keratin shows that this tissue is a highly ordered nanostructure of the appropriate size to produce the observed hue by constructive interference. The similarity in nanostructure between the medullary keratin of *C. maynana* and other structurally coloured feather barbs indicates that constructive interference probably also causes the production of the structural greens, blues, violets and ultraviolet colours in other feather barbs^{1–5}.

Richard O. Prum*†, Rodolfo H. Torres‡, Scott Williamson†, Jan Dyck§

*Natural History Museum, and Departments of

†Ecology and Evolutionary Biology, and

‡Mathematics, University of Kansas,

Lawrence, Kansas 66045, USA

e-mail: prum@ukans.edu

§Institute of Population Biology,

Universitetsparken 15,

DK-2100 Copenhagen, Denmark

1. Dyck, J. *Biol. Skrifter*, **18**, 1–67 (1971).

2. Dyck, J. *Z. Zellforsch.*, **115**, 17–29 (1971).

3. Dyck, J. *Proc. Int. Ornithol. Congr.*, **16**, 426–437 (1976).

4. Fox, D. L. *Animal Biochromes and Structural Colors* (Univ. California Press, Berkeley, 1976).

5. Durrer, H. in *Biology of the Integument* Vol. 2. *Vertebrates* (eds Bereiter-Hahn, J., Matoltsky, A. G. & Richards, K. S.) 239–247 (Springer, Berlin, 1986).

6. Finger, E. *Naturwissenschaften*, **82**, 570–573 (1995).

7. Benedek, G. B. *Appl. Optics*, **10**, 459–473 (1971).

8. Bohren, C. F. & Huffman, D. R. *Absorption and Scattering of Light by Small Particles* (Wiley, New York, 1983).

9. Gill, F. B. *Ornithology* (Freeman, New York, 1995).

10. Briggs, W. L. & Henson, V. E. *The DFT* (Society for Industrial and Applied Mathematics, Philadelphia, 1995).

11. *MATLAB Reference Guide* (Mathworks, Natick, MA, 1992).

12. Dyck, J. *Anser* (suppl.), **3**, 57–75 (1978).

A lower jaw from a Cretaceous parrot

All known Cretaceous bird fossils representing modern higher taxa are from the aquatic groups Anseriformes^{1–3}, Gaviiformes^{4,5}, Procellariiformes¹ and Charadriiformes^{1,6}. Here I describe a toothless avian dentary symphysis (fused jawbone) from the latest Cretaceous of Wyoming, United States. This symphysis appears to represent the oldest known parrot and is, to my knowledge, the first known fossil of a 'terrestrial' modern bird group from the Cretaceous. The existence of this fossil supports the hypothesis, based on molecular divergence data^{7,8}, that most or all of the major modern bird groups were present in the Cretaceous.

The specimen (Fig. 1) is from UCMF locality V65238 in the Lance Formation, Lancian North American Land Mammal Age, Maastrichtian, latest Cretaceous, Niobrara County, Wyoming. The mandibular symphysis is deeply concave, with a nearly U-shaped cross-section at the posterior end of the fragment, and has a rounded anterior margin (Fig. 1). The left and right dentaries are fused completely without a suture. Small bone fibres indicate that the individual was at or close to adult size, so the characters are not juvenile states. The labial surface of the mandibular symphysis has a highly vascularized appearance, produced by numerous foramina (openings), canals and grooves for neurovascular tissue (Fig. 1c), like those present in the area in which a horny covering, or rhamphotheca, is attached to the jaw in living amniotes.

A completely fused and edentulous (toothless) mandibular symphysis is present only in turtles, oviraptorids, ornithomimids, caenagnathid theropods and neornithine birds. The pair of lingual median symphyseal foramina next to the sagittal plane is found in most neornithines, including Anseriformes, Galliformes, Psittaciformes and other orders, but is absent in hesperornithiforms, oviraptorids and turtles. Caenagnathid theropods have lateral grooves, lateral occlusal grooves, lingual ridges, longitudinal vascular grooves, meckelian grooves and median tubercles⁹, which

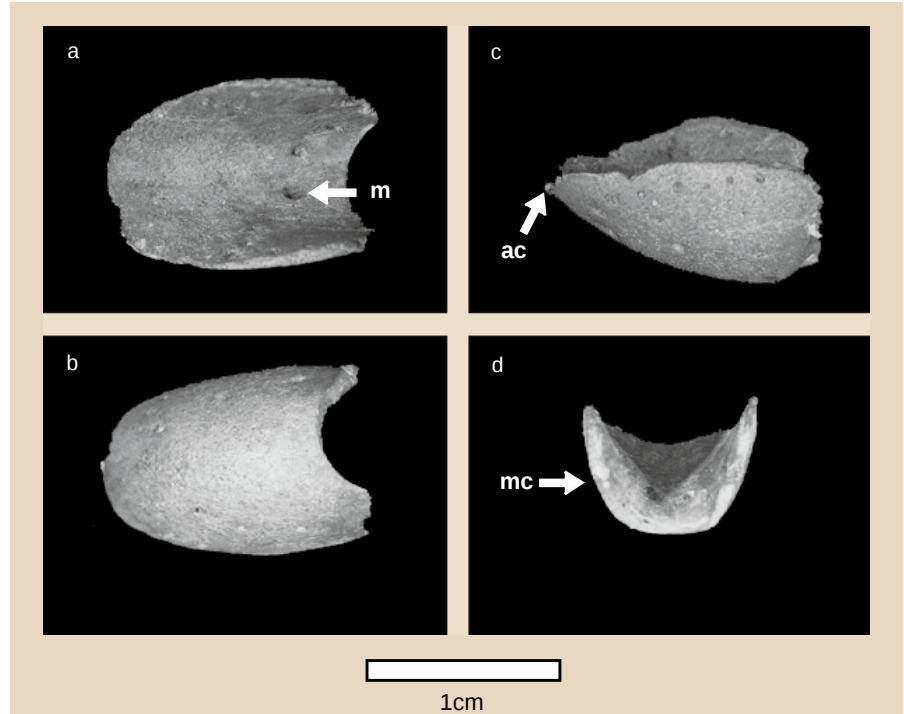


Figure 1 Mandibular symphysis of the Lance Formation parrot (University of California Museum of Paleontology, specimen UCMF 143274). a, Dorsal view. The two large median foramina (m) on the lingual surface 3.0 millimetres from the posteroventral edge of the symphysis are equidistant from the midline, with very short grooves posterior to them. b, Ventral view. c, Dorsolateral view, showing anterior accessory canal opening (ac) and large labial foramina; the right tomial crest is slightly concave, curving ventrally, and the left is worn and does not have a concave dorsal margin. d, Posterior view, showing meckelian canal (mc). Anterior is to the left in a–c.

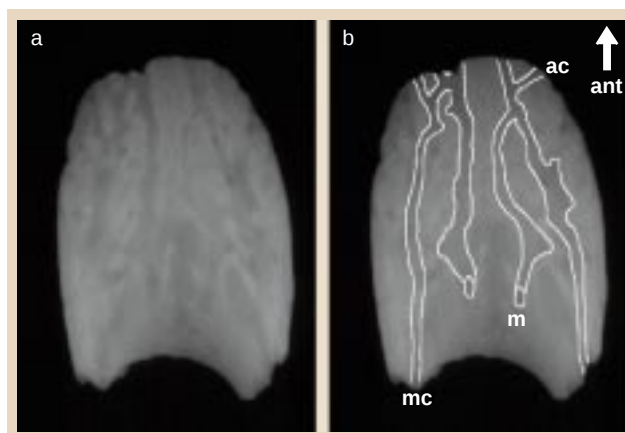


Figure 2 X-rays of the dorsoventral view of the Lance Formation parrot mandibular symphysis. **a**, The mandibular symphysis. **b**, The mandibular symphysis with the neurovascular canals outlined. The median neurovascular canals extend from the median foramina to the anterior end of the symphysis and intersect the meckelian canals. Accessory canals extend from the median canals, anterior to the intersection with the meckelian canals, anterolaterally to the tomial crest. This results in a K-shaped neurovascular canal pattern in the right side of the symphysis, with a mirror image of the pattern in the left side. Ant, anterior; other abbreviations are as in Fig. 1.

are absent in extant parrots and this specimen; the caenagnathid median symphyseal foramina are located anterior to the position seen in parrots and this fossil⁹. The K-shaped neurovascular canal pattern (Fig. 2) and deeply concave symphysis seen in this specimen are apparently derived characters found only in Psittaciformes. The rounded rostral end of the symphysis, the deeply concave symphysis (compared with most Psittaciformes) and the concave tomial crest are most common in Loriidae, but also occur in some macaws and other psittacids.

The discovery of this parrot in the Lance Formation indicates that the lineage leading to the parrot crown group was present by the end of the Cretaceous. If this parrot were a lory, as suggested by its morphology, the most recent common ancestor of the psittaciform crown group would be placed in the Cretaceous¹⁰, as supported by molecular data¹¹. The occurrence of a parrot in the Cretaceous implies the presence of other closely related bird taxa in the Cretaceous, as also predicted by molecular divergence data^{7,8}. These data also indicate that modern bird groups, including parrots, may have been relatively unaffected by the mass extinction at the end of the Cretaceous period.

Thomas A. Stidham

Department of Integrative Biology, Museum of Paleontology, and Museum of Vertebrate Zoology, University of California, Berkeley, Berkeley, California 94720, USA
e-mail: furcula@socrates.berkeley.edu

- Olson, S. L. & Parris, D. *Smithson. Contr. Paleobiol.* **63**, 1–22 (1987).
- Elzanowski, A. & Brett-Surman, M. K. *Auk* **112**, 762–767 (1996).
- Tambussi, C. P. & Noreiga, J. I. *Ameghiniana* **32**, 57–61 (1996).
- Chatterjee, S. J. *Vert. Paleontol. Abstr.* **9**, 16A (1989).
- Olson, S. L. *J. Vert. Paleontol.* **12**, 122–124 (1992).
- Brodtkorb, P. in *Proc. XIIIth Int. Ornith. Congr.* (ed. Sibley, C. G.) 55–70 (American Ornithologists Union, Ithaca, NY, 1963).
- Cooper, A. & Penny, D. *Science* **275**, 1109–1113 (1997).
- Hedges, S. B., Parker, P. H., Sibley, C. G. & Kumar, S. *Nature* **381**, 226–229 (1996).
- Currie, P. J., Godfrey, S. J. & Nesson, L. *Can. J. Earth Sci.* **30**, 2255–2272 (1993).
- Christidis, L., Schodde, R., Shaw, D. D. & Maynes, S. F. *Condor* **93**, 302–317 (1991).
- Miyaki, C. Y., Matiolli, S. R., Burke, T. & Wajntal, A. *Mol. Biol. Evol.* **15**, 544–551 (1998).

- Christidis, L., Schodde, R., Shaw, D. D. & Maynes, S. F. *Condor* **93**, 302–317 (1991).
- Miyaki, C. Y., Matiolli, S. R., Burke, T. & Wajntal, A. *Mol. Biol. Evol.* **15**, 544–551 (1998).

Hydrologic cycle explains the evaporation paradox

The evaporation of water, measured using evaporation pans, has been decreasing in the past few decades over large areas with different climates. The common interpretation is that the trend is related to increasing cloudiness, and that it provides an indication of decreasing potential evaporation and a decreasing terrestrial evaporation component in the hydrologic cycle. Here we show that, although these studies are valuable, pan evaporation has not been used correctly as an indicator of climate change.

At first glance, reports of decreasing pan evaporation in European Russia, Siberia and the western and eastern United States¹, India² and Venezuela³ are paradoxical. They are hard to reconcile with well-substantiated increases in global precipitation and cloudiness⁴, which would normally require more surface evaporation as the only source of atmospheric water vapour, rather than less. They also run counter to predictions of increasing evaporation⁵, as one of the more robust outcomes of radiative forcing, resulting from increasing atmospheric CO₂ in global circulation model calculations.

We resolve this paradox by demonstrating that, in non-humid environments, measured pan evaporation is not a good measure of potential evaporation; moreover, in many situations, decreasing pan evaporation actually provides a strong indication of increasing terrestrial evaporation.

The evaporation from a pan, E_{pa} , can be used as a good indicator of the evaporation, E , from the surrounding environment, but only when land-surface moisture is in

ample supply; this normally involves multiplication by a 'pan coefficient' a of order one, depending mainly on pan type⁶. The evaporation from any large uniform land surface, with adequate moisture so available energy is the limiting factor, is usually referred to as potential evaporation, E_0 . The actual evaporation from a well-watered surface is $E = E_0 = aE_{pa}$. Whenever the land-surface moisture becomes limiting and insufficient to sustain E_0 , the actual evaporation, E , decreases below E_0 and the energy not used up by E manifests itself as an increase in sensible heat flux ΔH , such that $E = E_0 - \Delta H$. This in turn causes aE_{pa} to exceed E_0 , or $aE_{pa} = E_0 + b\Delta H$, where b is another coefficient slightly larger than one, again depending mainly on pan type. Now aE_{pa} no longer provides a direct measure of E_0 , so it is more appropriately called 'apparent' potential evaporation.

The main point is that E and E_{pa} exhibit complementary rather than proportional behaviour; indeed, for instance in the extreme case of a desert environment, E is zero, whereas E_{pa} is at its maximum. The idea of a complementary relationship between actual evaporation and apparent potential evaporation is not new⁷, and it has stimulated advances in the estimation of terrestrial evaporation^{8–10}. In the case of a pan filled with water and placed in a region with less than adequate ground wetness to sustain E_0 , elimination of ΔH in the above yields $E = [(1 + b)E_0 - aE_{pa}]/b$.

Because a and b are of order one, this equation indicates how the observed^{1–3} decreases in pan evaporation, E_{pa} , can be interpreted as evidence for increasing terrestrial evaporation, E , in those regions. This is consistent with data^{4,11} indicating an intensifying hydrologic cycle in large regions: increasing precipitation leads to increasing surface run-off and soil wetness, which in turn generates more evaporation, and so on.

W. Brutsaert*, M. B. Parlange†

*School of Civil and Environmental Engineering, Cornell University, Ithaca, New York 14853, USA

†Department of Geography and Environmental Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA

- Peterson, T. C., Golubev, V. S. & Groisman, P. Y. *Nature* **377**, 687–688 (1995).
- Chattopadhyay, N. & Hulme, M. *Agric. Forest. Meteorol.* **87**, 55–73 (1997).
- Quintana-Gomez, R. A. *Proc. 7th International Meeting on Statistical Climatology*, 25–29 May 1997 (Whisler, BC, Canada, 1998).
- Karl, T. R., Knight, R. W., Easterling, D. R. & Quayle, R. G. *Bull. Am. Meteorol. Soc.* **77**, 279–292 (1996).
- Manabe, S. *Ambio* **26**, 47–51 (1997).
- Brutsaert, W. *Evaporation into the Atmosphere* (Kluwer Academic, Dordrecht, 1982).
- Bouchet, R. J. *Int. Assoc. Sci. Hydrol. Pub.* **62**, 134–142 (1963).
- Morton, F. I. J. *Hydraul. Div. ASCE* **102**, 275–291 (1976).
- Brutsaert, W. & Stricker, H. *Wat. Resour. Res.* **15**, 443–450 (1979).
- Parlange, M. B. & Katul, G. G. *Wat. Resour. Res.* **28**, 127–132 (1992).
- Lins, H. F. & Michaels, P. J. *Eos* **75**, 281, 284, 285 (1994).

Crude weapons of science's 'friends'

A House Built on Sand: Exposing Postmodernist Myths about Science

edited by Noretta Koertge

Oxford University Press: 1998. 322 pp.

£29.95, \$35

Jonathan Rée

A generation ago, scientists who got depressed by the pettiness of everyday professional life could seek comfort in the flattering images of science that circulated outside their laboratories. In the science fiction of H. G. Wells or Arthur C. Clarke science was the road to technotopia; in the historical sociology of Robert K. Merton and J. D. Bernal it was the engine of social progress; and in the intellectual history of Oliver Lodge or J. G. Crowther it was one of the glories of human intellect and imagination.

But not any more. Science fiction has gone out of fashion, at least in its optimistic forms, and historians and sociologists have created a new academic field — science studies — in which science is more likely to get buried than praised. Scientists and their supporters are understandably upset, and some of them have started to retaliate in the name of an “objective truth” that they take to be under attack from “postmodern relativism”.

In 1994 the molecular biologist Paul Gross and mathematician Norman Levitt sounded the alarm in *Higher Superstition: The Academic Left and Its Quarrels with Science*, followed a year later by a collection called *The Flight from Science and Reason*. Panic increased in 1996 when the physicist Alan Sokal hoaxed the journal *Social Text* into printing a satirical spoof called “Transgressing the boundaries: toward a transformative hermeneutics of quantum gravity” as if it were a serious piece of research. Sokal then prolonged his joke into a book called *Intellectual Impostures*, and now, in *A House Built on Sand*, Noretta Koertge has brought together 16 worried authors, including Gross, Levitt and Sokal, for a concerted counterattack on those they regard as dangerous enemies of modern science.

It is unlucky for Koertge's troops that their indictment comes out at the same time as Kenneth Starr's report on the misdemeanours of the US President. No one can deny that they have discovered a few pieces of shoddy scholarship masquerading as radical or feminist science studies: an unfortunate paper on “Gender encoding in fluid mechanics”, for instance, and an ill-considered journalistic tirade about “bashful eggs and macho sperm”, and perhaps a slippery slide between Einsteinian relativity and Pro-



MARK DOBSON

tagorean relativism. But officious indictments make dull and unedifying reading, and many readers will find it easier to identify with the accused than with the prosecutors. After all, no field of research is completely free of coat-trailing, sycophancy, dullness, inaccuracy, stupidity, distortion and misreporting. Is it not nobler to ignore such lapses rather than seek them out and gloat?

Moreover there is a joker in Koertge's pack. The jacket blurb promises that Philip Kitcher will “expose the relativist epistemology that underlies postmodernist accounts of science”, but his essay is way off message. His “plea for science studies” reminds us how difficult it is for anyone to keep their balance in a discipline that is constituted by two conflicting truisms: first, that science progressively enhances our powers of prediction and control, and second, that science is part of the natural history of fallible humanity. It is not surprising if the attempt sometimes fails,

and fails badly; but Kitcher is able to describe numerous successes as well, citing Nancy Cartwright, Lorraine Daston, Peter Galison, Ian Hacking, Martin Rudwick and many others. He then notes that these fine authors are systematically ignored by the self-appointed friends of science, who prefer to recycle a small number of egregious examples of postmodern vice — cases which, Kitcher says, “practitioners of science studies would not view as central to the field”, and which he himself had “never heard of” before the scientists started using them as evidence for the prosecution.

There is a fearful symmetry between the two sides in this debate. Proponents of the cheaper kind of science studies like to collect embarrassing stories about scientists who have been arbitrary, dishonest, biased or hasty, and they love to expose venal motives at work in the heartlands of science. Meantime the alarmist friends of science enjoy collecting embarrassing stories about

postmodernists who have been arbitrary, dishonest, biased or hasty, and they are never happier than when they can discover venal motives in the heartlands of science studies.

And the difficulties facing the two sides mirror each other too. The postmodernists find it hard to see why science still makes progress, despite all its follies and vicissitudes, or why individuals, firms and governments find it profitable to invest their time and money in it. On the other hand the friends of science have difficulty accounting for the fact that science requires perspiration and inspiration as well as passive observation; indeed they seem incapable of explaining why science should have a long and laborious history at all, least of all a history of detours, disagreements and dead ends — a history which might well, for all we know, have ended up somewhere completely different.

But the issues that divide science from science studies are never going to be resolved by counting up malpractices or unsolved problems on each side. The fact that researchers make pratfalls in the course of their enquiries does not discredit their intellectual ambitions. And if one discipline is more chaotic or accident-prone than another, that may only indicate that its requirements are particularly rigorous. The adjudication of such disputes calls for direct theoretical arguments rather than accumulations of insults and anecdotes.

The authors of *A House Built on Sand* think they have such an argument, for they picture themselves as defenders of “objective truth” against “relativism”. But it is possible that they are deceived. They spend a lot of time chastising amateurish commentators who sound off about scientific issues without mastering the relevant literature, but they court exactly the same criticism when they pontificate about epistemology and conceptual schemes without showing the slightest acquaintance with Carnap, Goodman, Davidson, Putnam, MacIntyre or McDowell, and without even tipping their cap towards Frege and Wittgenstein or Bachelard or Canguilhem, or Kant, Hegel, Husserl, Heidegger and Habermas. Someone should have told them that philosophy, like quantum mechanics, contains tricky terrains where angels fear to tread.

Koertge informs us that the language used by scientists is “characterised by clarity, precision, and economy”, whereas relativists, she thinks, cannot even show elementary respect for “literal meaning”. But pride comes before a fall, and the phrase “objective truth” may be more of a banana skin than she and her colleagues realize. Sometimes they seem to think it describes the distinguishing feature of science as such, and that truth becomes objective by being grounded in experiments that map out physical mechanisms behind the surface facts of nature. But

such a definition would leave half of science out in the cold, since the truths of theoretical linguistics or artificial intelligence or number theory are certainly not maps of physical reality.

On other occasions Koertge’s authors use “objective truth” to hint at an implicit category of “subjective truth”, to which they would consign everything that is not science. But this will not serve their purposes either. If “subjective truth” is more than just a figurative description of falsehood, it must refer either to statements that are purely formal, like those of logic, or to statements describing emotions or personal attitudes of mind. But when such statements are true, their truth is as objective as anyone could desire. And when they are false, their failing is not that they are subjective, but that they are not true.

Paul Gross gets very irritated by feminists who, as he puts it, try to “redefine all meaning out of ‘objectivity’”. But it may be that they have put their finger on a sensitive point, for what exactly can he mean by the “objectivity” on which he stakes his professional honour? We can hardly add anything to the stature of truth by awarding it the epithet “objective”: what other kind of truth is there supposed to be? The term may have had precise meanings in medieval treatises on logic and metaphysics, but in the dialect of today’s friends of science “objective truth” functions rather like “God’s truth” or “the honest truth” — an emotive intensifier, without any clear or literal meaning.

The word “relativism” may prove an equally slippery accomplice. There would be no point in writing long and bad-tempered protests against relativism if it were merely the doctrine that there is no such thing as knowledge: we could dismiss it instantly by asking the question, How do you know? But we can also make relativism into something passably intelligent, namely an attempt to explain how we can have reliable knowledge, given that our thinking is shaped, willy nilly, by the particular languages we happen to speak, or the taxonomies and measuring systems we have inherited along with them. The goal of intelligent relativism would be to stop us daydreaming about a utopia of immaculate knowledge, and persuade us to settle for such partial certainties as the human condition allows. Clearly relativism in this sense is a gentle encouragement to scientific endeavor.

The issues that divide science from science studies are never going to be resolved by counting up malpractices or unsolved problems on each side

our, not a counsel of epistemological despair.

The most substantial section of *A House Built on Sand* is Meera Nanda’s account of the “People’s Science Movements” which were launched in India in the 1960s. They aimed to diffuse a basic understanding of science (the germ theory of disease for instance, or the laws of Newtonian physics) so as to disseminate a “critical scientific temper that would subject traditional knowledge claims to the test of empirical evidence”. By the 1980s the People’s Science Movements involved some 20,000 activists, including Nanda herself. But, she tells us, they are now reeling under the combined assault of right-wing Hindu fundamentalists and left-wing secular anthropologists, who have formed a strange alliance in defence of the “epistemological rights” of traditional knowers — that is to say sorcerers, goddess-worshippers and widow-burners, whose beliefs are liable to be annihilated by an unrestricted spread of “Western” science. Nanda graciously declines “the privilege of having our traditional knowledge considered at par with science”, noting that such “epistemic charity” is really the cruellest form of condescension. “The project of different and equal sciences for different people”, as she puts it, “completely negates our project of science for all people.”

But the vividness of Nanda’s description of “epistemic charity” is not matched by a clear philosophical analysis. Like the other contributors to *A House Built on Sand*, she lays the blame for “antiscience” on a conspiracy of “postmodern relativists” who insist on regarding our beliefs as “socially constructed” and refuse to make any exception for science or “objective truth”. A different diagnosis is possible, however. In the first place, the Indian traditionalists who are hostile to science are not themselves relativists: given the choice, they would regard their own beliefs as “objectively true”, just like the modern friends of science. If we want to convert them to a different way of thinking, then relativism may be precisely what we need: for if both science and superstition are regarded as social constructions, we can simply invite the traditionalists to give science a try, in the firm expectation that they will find it suits them better.

On the other hand, the flaw in the sinister doctrine of “epistemic charity” is not so much its relativist theory of knowledge as its absolutist conception of “cultures”. In fact, the advocates of epistemic charity are not really relativists at all: if they were, they would not be trying to jump out of their own cultural skins and treat different human “cultures” with high-altitude impartial objectivity. Consistent relativists would realize that our cultures — whether scientific or antiscientific or somewhere in between —

are not impregnable island fortresses entire unto themselves, but networks of communication: they reach out to the world and ask to be transformed by it.

Scientists have no reason to panic at the thought that science is an element in human culture alongside many others. Science is not about to be marginalized or engulfed, as European magic and witchcraft were 300 years ago. Scientists may exasperate their cultural neighbours by constantly shouting abuse at “relativists” and beating on the drum of “objective truth”, but on the whole they still have the public’s approval. What should really worry us is that science is still being used — perhaps now more than ever — as a means of cultural exclusion, dividing the world into scientific haves and have-nots. It urgently needs to make itself more welcoming to strangers, so as to become a part not only of our own culture, but of culture “for all people”. Sanctimonious, defensive sarcasm can only make matters worse. □

Jonathan Rée is in the Department of Philosophy, Middlesex University, White Hart Lane, London N17 8HR, UK.

Wondrous correctness

Unweaving the Rainbow
by Richard Dawkins
Penguin: 1998. 313 pp. £20, \$26
Gillian Beer

This volume falls, somewhat prematurely we may hope, into the Grand Old Man genre. Richard Dawkins lards his enthusiastic tribute to the wonders of science with quirky asides and personal anecdotes. These anecdotes assume the reader’s affection for him, always a risky assumption before reaching admired and advanced old age. His conversational style allows him to be brusque and testy about others’ achievements and expect to carry the reader with him.

The argument of the book is unobjectionable: that understanding the processes revealed by scientific analysis makes things more wonderful, not less. Emphasizing the Wonders of Science has been a staple of scientific popularization for the past 200 years. Science makes strange the familiar and thus opens our eyes to the intricacy and the extent of the world within which we bumble along from day to day. Dawkins makes a further turn on this argument to distinguish between ‘good’ and ‘bad’ wonder. He attacks credulity and its manifestations “in superstition, the paranormal, and astrology”.

John Tyndall undertook a similar epistemological joust, more amusingly described; invited to attend a seance 100 years ago, and fresh from his Royal Institution demonstrations, he competed with his spiritualist

hosts: “The wonderful narratives resumed; but I had narratives of my own quite as wonderful. These spirits, indeed, seemed clumsy creations, compared with those with which my own work had made me familiar. I therefore began to match the wonders related to me with other wonders. A lady present discoursed on spiritual atmospheres, which she could see as beautiful colours when she closed her eyes. I professed myself able to see similar colours, and, more than that, to be able to see the interior of my own eyes. The medium spoke of the performances of the spirits on musical instruments. I said that such a performance was gross, in comparison with a kind of music that had been discovered some time previously by a scientific man. Standing at a distance of twenty feet from a jet of gas, he could command the flame to emit a melodious note.... These were acknowledged to be as great marvels as any of those of spiritdom. The spirits were then consulted, and I was pronounced to be a first-class medium.”

Tyndall recognizes with wry humour how readily scientific assertion can be absorbed back into contrary beliefs. Dawkins takes that argument in a different direction. He enlarges his attack on credulity to include what he sees as falsely ethical readings of evolution in the work of Stephen Jay Gould and in followers of James Lovelock. He offers, as an alternative, clear and approving accounts of the work of scientists such as Horace Barlow and Richard Gregory, who have paid special attention to vision and its determinants.

The book is at its best in those chapters and passages where Dawkins can delight in exposition; he writes effectively on genes and their histories, and on the brain and its modelling powers. Sometimes, however, the chapters seem to be patched together from paragraphs and *aperçus* laid end to end without a defining thread of argument. Dawkins seems at times uncertain of the audience he is addressing (we are assumed to be knowing enough to share his prejudices against quite sophisticated critical theorists and yet to need educating about coincidences). Sometimes he denounces metaphor, and sometimes breezily adopts it. He has clearly been stung by the critiques of his own metaphoric habits in *The Selfish Gene* with its shift of levels from activity to ethics, for he alludes more than once to the need to read the whole book and not be misled by its title.

Unweaving the Rainbow is at its worst in the often impatient and cavalier treatment of evidence from intellectual fields outside science. This ranges from a demeaning reference to John Ruskin, and a simplistic aside concerning the cultural anthropologists Margaret Mead and Derek Freeman, to a condescending habit of wresting a line from a poem to serve his purpose, as if it had no further complexity or context. So, a fine quo-



Lamia: woman or serpent? The fruits of scientific enquiry may not always be pleasing to the eye.

tation from the astrophysicist Subrahmanyan Chandrasekhar is pitted against “Beauty is Truth, Truth Beauty” as sounding “much more sincere”.

Dawkins accuses his avowedly favourite poets, Keats and Yeats, of typifying an ignorant repudiation of science. If only Keats had turned to Sir Isaac Newton for inspiration, if only Yeats had accorded more value to reason, how much better their poetry would be! The title of his study draws on Keats’s poem “Lamia”, which is concerned with the peculiarly equivocal appearances of things. Lamia is a woman but she is also a destructive serpent; in both incarnations, however, she is very beautiful. As a serpent she is:

*A gordian shape of dazzling hue,
Vermilion-spotted, golden, green and blue,
Striped like a zebra, freckled like a pard,
Eyed like a peacock, and all crimson barr’d*

Which is her true nature? Is she, as she claims, an innocent woman bewitched, or is she a guileful serpent disguised? The old sophist Apollonius blasts her with his philosophic gaze and she resolves into a serpent, but she is no longer the beauty of before, now she withers away.

The vigour of Keats’s language thrives on precise detail. The poem struggles, with poignant sophistication, to interpret the cost of pursuing knowledge. It works to disabuse the reader from any idealized fancy that beauty will always be rediscovered at the end of enquiry. Keats had begun training as a

medical student and understood these issues without sentimentality.

Dawkins wants it both ways. He wants to function as Apollonius, disabusing his reader of the various magical-seeming possibilities of astrology, coincidence, relativism and misleading metaphor. But he also wants to assert the inevitably wonder-enhancing power of scientific insight; he does not want to destroy the beauty of Lamia. He is dismayed by the poet's question:

*Do not all charms fly
At the mere touch of cold philosophy?
There was an awful rainbow once in heaven:
We know her woof, her texture; she is given
In the dull catalogue of common things*

In "Lamia" the question is loaded with the dismay of the poet struggling to locate truth. "Lamia" is no ignorant rejection of natural philosophy's reductionism; it is a painful and sinewy debate, more tough-minded than the softened prose in which Dawkins ends *Unweaving the Rainbow*: "A Keats and a Newton, listening to each other, might hear the galaxies sing". Or, as Keats recognized, the galaxies may not be singing. □

Gillian Beer is at Clare Hall, University of Cambridge, Herschel Road, Cambridge CB3 9AL, UK.

The unread menace

Measuring Minds: Henry Herbert Goddard and the Origins of American Intelligence Testing
edited by Leila Zenderland
Cambridge University Press: 1998. 448 pp.
£45, \$64.95

Roy Porter

Henry Herbert Goddard won a somewhat dubious immortality by coming up with the term 'moron' — meant to identify from among the mass of the 'feeble-minded' a "person of attested mental development with an intelligence comparable to that of the normal child between 8 and 12 years inclusive". From such a coinage, there is much that could be easily — and correctly — predicted about the convictions and career of one of the most influential US psychologists of the first half of this century (he died in 1957 at the ripe old age of 91). Goddard, as one might guess, was one of the vociferous cohort of scientists preoccupied with human 'degeneration' and seeking ways of averting the 'threat' posed by 'defectives' to American society.

But, as Leila Zenderland demonstrates in a well-researched if somewhat hefty biography, Goddard, although a keen eugenicist, resists being reduced to the reactionary eugenicist villain who looms large in recent historiography. After all, he thought of his politics as being progressive — he even favoured the New Deal — and he was

opposed to the more extreme measures (such as the 'lethal chamber') being touted to deal with the 'menace', and even rather doubtful about the desirability of surgical sterilization.

Goddard, who around the time of the First World War was America's most widely-read psychologist, saw his career in terms of chance and happenstance, and there is an element of truth in that view. Born in 1866 into a typically pious New England Quaker family, he attended a minor college (Haverford) and drifted around for a while on the fringes of higher education until, like so many others of his generation, he was fired by G. Stanley Hall of Clark University in Massachusetts. It was Hall who persuaded Freud to make his one and only trip to the United States. In Hall's inspiring vision, the up-and-coming discipline of evolutionary psychology was destined, in effect, to replace religion. It would bestow on society a body of scientific values suitable for combating the ills of modern times and grant to individuals a means to self-awareness and self-improvement. Goddard was converted.

Convinced that the 'unfit' constituted not only a pressing problem but a group on whom he could practise his new-found professional expertise, Goddard took employment in 1906 in a 'training school' for the 'feeble-minded' — Vineland, in rural New Jersey. Two years later, while travelling in Europe, he had the good fortune to be one of the first to grasp the significance of the intelligence tests newly developed by Alfred Binet, which launched the concept of 'mental age' within the framework of a general mental test.



The 'feeble-minded' underclass: these 'cases' of Goddard's were said to have a mental age of seven.

Confusion had long reigned when it came to the tricky matter of classifying the different grades of 'idiots', to say nothing of disputes — between doctors, educationists and institutional superintendents — as to the cause and proper treatment of the condition. An energetic member of the 'Feeble-Minded Club', Goddard was able to persuade first himself and then his colleagues that the Binet test — that is, a psychological approach — would provide solutions as to what should be done with these problem people.

Like many of his ilk who were, by training, pedagogues rather than physicians, Goddard was initially optimistic about what could be achieved: the 'feeble-minded' would prove educable, leading to "a better mind if not a perfect mind", if only one researched the right way to go about it. But bitter experience seemed to prove the opposite. And so his thinking underwent a sea-change, from a position not unsympathetic to the contribution of the environment ('nurture') in producing and rectifying defectives, to one that insisted on the cardinal role played by heredity ('nature').

Familiarizing himself with Mendelian genetics, Goddard began to investigate the family backgrounds of institutional populations and professed surprise on finding a vast submerged iceberg of imbecilism. Degeneracy evidently ran in families, and marriages between epileptics, alcoholics, criminals, syphilitics, nymphomaniacs and all other "defectives, dependants and delinquents" only served to make bad worse down the generations. The defectives were not only different; they might even constitute a 'moron majority'.

These findings were written up in 1912 in a book that became a bestseller: *The Kallikak Family: A Study in the Heredity of Feeble-Mindedness*. The name was fictitious (it is a compound of the Greek words for beautiful and bad), but the family was real. Goddard had traced its bifurcated pedigree (some 480 descendants) from its eighteenth-century roots. One branch was thoroughly respectable, while the other (descended from the first Kallikak's illegitimate offspring) engendered a succession of defectives, criminals, prostitutes and so forth. The book entered the culture, and it even achieved a reprint in Nazi Germany. A more conventional statement of his research findings then appeared in *Feeble-Mindedness: Its Causes and Consequences*, published in 1914.

The rest of Goddard's lengthy career was dedicated to further exposure of such undesirables and to the application of psychometric testing to all walks of life — from immigration controls to the army — in the belief that intelligence differentials were crucial to the understanding and resolution of social problems.

In Goddard's eyes, 'the facts' had thus

forced him to become a eugenicist. Experience equally seemed to suggest what should be done. Sterilization was, he thought, a suspect option — after all, civil liberties were central to American values. The way forward must lie in the institutional segregation of the unfit. Not only would that prevent defectives from breeding and create a supportive and humane environment for them, but it would provide a superb “human laboratory” (Goddard’s standard phrase) for researching their mentalities and laying bare the pathology of the human psyche.

It would, as Zenderland persuasively argues, be misleading to cast Goddard simply as some sort of stock bigot. Doubtless he believed there was some kind of underclass, but he was remarkably free of racial and colour prejudice — what he mainly feared were poor whites. He is best seen primarily as a representative of an emergent cadre of experts, scientists and professional administrators, anxious to establish a place in the sun for themselves as the new priesthood serving a secularizing society, preaching the gospel not of laissez-faire capitalism but of informed social responsibility.

Zenderland does not pretend that her protagonist was a very profound or original thinker. Although a passionate champion of ubiquitous intelligence testing, Goddard never seems to have thought deeply about what precisely it was that was being measured. He was a doer, a technician, lucky enough to hold in his hands, in the Binet test, that device utterly appropriate to the needs of classification and control in a mass society. □

Roy Porter is at the Wellcome Institute for the History of Medicine, 183 Euston Road, London NW1 2BE, UK.

Not black and white

Melanism: Evolution in Action

by Michael E. N. Majerus

Oxford University Press: 1998. 338 pp. £55, \$105 (hbk), £23.95, \$45 (pbk)

Jerry A. Coyne

From time to time, evolutionists re-examine a classic experimental study and find, to their horror, that it is flawed or downright wrong. We no longer use chromosomal polymorphism in *Drosophila pseudoobscura* to demonstrate heterozygous advantage, flower-colour variation in *Linanthus parryae* to illustrate random genetic drift, or the viceroy and monarch butterflies to exemplify Batesian mimicry. Until now, however, the prize horse in our stable of examples has been the evolution of ‘industrial melanism’ in the peppered moth, *Biston betularia*, presented by most teachers and textbooks as the paradigm of natural selection and evolution occurring within a human lifetime. The re-examination of this tale is the centrepiece of



Cautionary tale: the classic account of industrial melanism in the peppered moth now looks flawed.

Michael Majerus’s book, *Melanism: Evolution in Action*. Depressingly, Majerus shows that this classic example is in bad shape, and, while not yet ready for the glue factory, needs serious attention.

According to the standard textbook litany, before the mid-nineteenth century, all *B. betularia* in England were white moths peppered with black spots, a form called *typica*. Between 1850 and 1920, *typica* was largely replaced by a pure black form (*carbonaria*) produced by a single dominant allele, the frequency of which rose to nearly 100% in some areas. After 1950, this trend reversed, making *carbonaria* rare and *typica* again common. These persistent and directional changes implied natural selection. In a series of studies, this conclusion was verified by several investigators, most prominently Bernard Kettlewell of Oxford.

According to these workers, the evolution of colour was caused by birds eating the moths most conspicuous on their normal resting site — tree trunks. The increase in black moths was attributed to pollution accompanying the rise of heavy industry. A combination of soot and acid rain darkened trees by first killing the lichens that festooned them and then blackening the naked trunks. The *typica* form, previously camouflaged on lichens, thus became conspicuous and heavily preyed, while the less visible *carbonaria* enjoyed protection and increased in frequency. After the passage of the Clean Air Acts in the 1950s, trees regained their former appearance, reversing the selective advantage of the morphs. This conclusion was bolstered by a geographical correlation between pollution levels and morph frequencies (*carbonaria* was most common in industrial areas), and most prominently by Kettlewell’s famous experiments which showed that, after releasing *typica* and *carbonaria* in both polluted and unpolluted woods, researchers recaptured many more of the cryptic than of the conspicuous form. The differential predation was supported by direct observation of birds eating moths placed on trees. Finally, Kettlewell demonstrated in the laboratory that each form had a behavioural

preference to settle on backgrounds that matched its colour.

Criticisms of this story have circulated in samizdat for several years, but Majerus summarizes them for the first time in print in an absorbing two-chapter critique (coincidentally, a similar analysis [Sargent *et al.*, *Evol. Biol.* **30**, 299–322; 1998] has just appeared). Majerus notes that the most serious problem is that *B. betularia* probably does not rest on tree trunks — exactly two moths have been seen in such a position in more than 40 years of intensive search. The natural resting spots are, in fact, a mystery. This alone invalidates Kettlewell’s release–recapture experiments, as moths were released by placing them directly onto tree trunks, where they are highly visible to bird predators. (Kettlewell also released his moths during the day, while they normally choose resting places at night.) The story is further eroded by noting that the resurgence of *typica* occurred well before lichens recolonized the polluted trees, and that a parallel increase and decrease of the melanic form also occurred in industrial areas of the United States, where there was no change in the abundance of the lichens that supposedly play such an important role.

Finally, the results of Kettlewell’s behavioural experiments were not replicated in later studies: moths have no tendency to choose matching backgrounds. Majerus finds many other flaws in the work, but they are too numerous to list here. I unearthed additional problems when, embarrassed at having taught the standard *Biston* story for years, I read Kettlewell’s papers for the first time.

Majerus concludes, reasonably, that all we can deduce from this story is that it is a case of rapid evolution, probably involving pollution and bird predation. I would, however, replace “probably” with “perhaps”. *B. betularia* shows the footprint of natural selection, but we have not yet seen the feet. Majerus finds some solace in his analysis, claiming that the true story is likely to be more complex and therefore more interesting, but one senses that he is making a virtue of necessity. My own reaction resembles the dismay attending my discovery, at the age of six, that it was my father and not Santa who brought the presents on Christmas Eve.

Occupying a quarter of the book, the *Biston* analysis is necessary reading for all evolutionists, as are the introductory chapters on the nature of melanism, its distribution among animals, and its proposed causes. Majerus, however, designed his book for both professional and lay readers, and this causes some unevenness in the material. The *Biston* story is sandwiched between less compelling chapters, including long sections on the basic principles of genetics and evolution, which can be skipped by evolutionists. Other discussions, involving melanism in ladybirds and other Lepidoptera, as well as

the author's unpublished work on habitat selection, are full of technical details that will overwhelm the lay reader. Unfortunately, most of the work described is inconclusive; despite the widespread occurrence of melanism, its evolutionary significance is nearly always unknown.

What can one make of all this? Majerus concludes with the usual call for more research, but several lessons are already at hand. First, for the time being we must discard *Biston* as a well-understood example of natural selection in action, although it is clearly a case of evolution. There are many studies more appropriate for use in the classroom, including the classic work of Peter and Rosemary Grant on beak-size evolution in Galapagos finches. It is also worth pondering why there has been general and unquestioned acceptance of Kettlewell's work. Perhaps such powerful stories discourage close scrutiny. Moreover, in evolutionary biology there is little payoff in repeating other people's experiments, and, unlike molecular biology, our field is not self-correcting because few studies depend on the accuracy of earlier ones. Finally, teachers such as myself often neglect original papers in favour of shorter textbook summaries, which bleach the blemishes from complicated experiments.

It is clear that, as with most other work in evolutionary biology, understanding selection in *Biston* will require much more information about the animal's habits. Evolutionists may bridle at such a conclusion, because ecological data are very hard to gather. Nevertheless, there is no other way to unravel the forces changing a character. We must stop pretending that we understand the course of natural selection as soon as we have calculated the relative fitness of different traits. □

Jerry A. Coyne is in the Department of Ecology and Evolution, University of Chicago, 1101 E. 57 Street, Chicago, Illinois 60637, USA.

More than meets the eye

Visual Intelligence: How We Create What We See

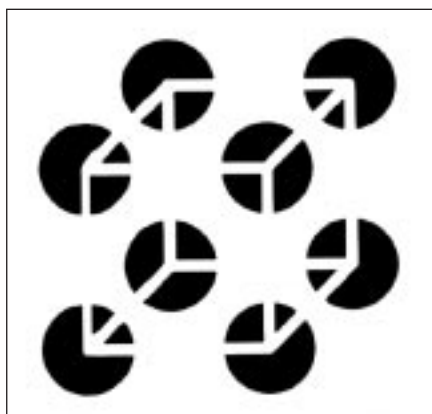
by Donald D. Hoffman

Norton: 1998. 294 pp. \$29.95, £21

John C. Marshall

It sometimes seems that a new visual area is discovered in the primate brain every week. The sheer amount of brain that has purportedly been colonized for vision makes one wonder how we can do anything other than see. It is likewise unclear whether this aggressive drive for *lebensraum* characterizes the biological evolution of the visual system or the cultural evolution of neuroscientists.

Be that as it may, it is quite a relief to see



The "subjective" Necker cube: the black discs can be seen as behind the cube or as holes in front of it.

that Donald Hoffman's riveting introduction to *Visual Intelligence* contains very little about brains. Even when he describes how brain damage can cause loss of colour perception or of motion perception, Hoffman has his eye more on the functional ramifications of the loss than on the neuroanatomical underpinnings. For the most part, then, Hoffman adopts the classical approach of cognitive science and computer science: describe the rules according to which the visual system operates, and let somebody else worry about which neurons are where.

But what is vision? Here Hoffman quotes (and follows) the position originally expounded by David Marr: "Vision is a process that produces from images of the external world a description that is useful to the viewer and not cluttered with irrelevant information . . ." And what in Hoffman's own terms is the fundamental problem of vision? That "the image at the eye has countless possible interpretations". And within which overall framework can the problem be solved? Hoffman sees a strong parallel between Noam Chomsky's arguments for rules of universal grammar and his own rules of universal vision. In both cases, the mature competence that quickly develops is grossly underdetermined by the fragmentary data presented to the senses.

Hoffman accordingly conjectures that "the innate rules of universal vision are part of the child's biology, and allow the child to acquire, through visual experiences that might vary from one culture to another, the rules of visual processing. The rules of visual processing, in turn, allow the visually competent child or adult to construct specific visual scenes by looking." Hoffman is happy to agree with James Gibson that retinal images, above all moving images, are "rich in information". They are just nowhere near rich enough to pick out our visual world from all the "countless possible visual worlds" that are compatible with such images.

Chapter two ("Inflating an artist's sketch") shows most clearly how Hoffman's

strategy operates. The issue tackled is a fundamental puzzle in depth perception: "The image at the eye has two dimensions; therefore it has countless interpretations in three dimensions." Hoffman then solves (most of) the problem of how the visual system comes to the correct interpretation (most of the time) by conjecturing an ordered sequence of visual rules. These range from the simple "Always interpret a straight line in an image as a straight line in 3D" to the considerably less obvious "Interpret each concave point on a bound as a saddle point on a rim".

Hoffman leads the reader through the justification for these maxims by showing with line drawings and other two-dimensional patterns exactly how each rule serves to constrain the percept that we actually derive from the image. In other chapters, Hoffman deploys essentially the same strategy to show how the visual system recovers surfaces, shapes and their parts, colour, and the path of moving objects by "an intelligent process of active construction".

Hoffman's book has many virtues, of which sheer intellectual excitement is the foremost. *Visual Intelligence* has been aimed at the lay reader ("tourists", as Hoffman calls them) and is indeed sufficiently lucid to attract and hold such an audience without insulting their intelligence. Each of the many figures illustrates an argument. And, as so often happens in the theory of perception, Hoffman can show that artists often had an intuitive understanding of principles that the scientists later 'discovered'. From Brunelleschi to Picasso by way of Dürer, painters were using pictorial devices that are only now coming to be formally understood. Hoffman even manages to find palaeolithic cave drawings of bison that show the kinds of subjective contours studied to such great effect by Gaetano Kanizsa.

But this is no 'coffee table book', and I would be surprised if even the most experienced of visual scientists do not learn much from Hoffman's guidance. Even more rare in a book of 'popular science', Hoffman acknowledges the sources of ideas and findings that are not his own work, and gives a full and accurate list of references.

The scattered citations of earlier philosophical understanding (or misunderstanding) of vision include highly appropriate discussions of Berkeley, Locke and Malebranche, although to my considerable surprise Plato's cave has gone missing. If there is one image that sums up the thrust of Hoffman's work, surely it is that dungeon. The moral of *Visual Intelligence* is that we have spent so long in the cave that our brains can now derive what is really out there from the merest flickering of shadows on the wall. □
John C. Marshall is at the Neuropsychology Unit, University Department of Clinical Neurology, Radcliffe Infirmary, Woodstock Road, Oxford OX2 6HE, UK.

Footing the bill

Gout: The Patrician Malady

by Roy Porter and G. S. Rousseau

Yale University Press: 1998. 393 pp. \$35, £25

Anne Crowther

Gout is a great subject for showing how the perception of illness is determined by social preoccupations. Most recently, the public's preoccupation with AIDS as a 'gay plague' reduced awareness of other types of transmission, just as in the past there were stereotypes of the 'consumptive' personality. Gout, although it is a widespread malady which causes great pain and can only be relieved, not cured, is still associated with a social profile developed most vigorously during the eighteenth century. It is an ailment chiefly associated with older males, and, by long-standing convention, with the rich.

Although Roy Porter and G. S. Rousseau note the current expansion of gout to countries where it was once little known, as a 'disease of civilization', they are less concerned with its modern position than its cultural heritage. Their book therefore mainly examines the work of medical theorists and the often detailed introspections of patients. Gout has a tremendous cast of characters: on the arts side there are Samuel Johnson, William Cowper, Edward Gibbon, Horace Walpole, Tobias Smollett and many more. On the science side are Francis Bacon and William Harvey (possibly), Thomas Sydenham, Benjamin Franklin and generations of Darwins.

The Greeks had been concerned about gout, though they did not distinguish it from arthritic problems. 'Podagra' was a suitably vague description, placing gout in its usual location, the foot, but the ailment was also suspected of mounting internal attacks, sometimes described as 'flying' or 'wandering' gout. The Greeks saw gout, like many other diseases, as due to an internal imbalance in the body, although the question of external influences was often raised.

Sydenham's *Treatise on the Gout* (1683), set the seal on gout's long-term definition as a disease of the wealthier classes. He also placed it as an imbalance of the constitution rather than an infectious or externally conditioned disease, holding that gout was related to luxurious living and overindulgence in alcohol. The logic of this type of analysis was that the body reacted to the morbid humours within itself by banishing them to the extremities, where they could do relatively little harm. Gout was not a fatal malady; indeed, it was believed to be a sign that the body was insulating itself against more serious illness. Nervous and sensitive temperaments, as many creative patients agreed, were likely to be most susceptible. They could also use their malady as an



Rich man's illness: gout has long been associated with the leisured classes, as seen here by Hogarth.

excuse for retreating into their sedentary work.

In 1771, William Cadogan's influential book on gout claimed that it was preventable and not hereditary, and that its predisposing causes were "excess and idleness". Porter and Rousseau show that the current language of gout reflected the political theories of the day; the same language was applied to disorders of the body and the body politic. The higher-class theorists offered impeccably logical explanations but not much hope of a cure. Sydenham recommended opium and digestives, Cadogan concentrated on regimen. But this did not prevent gout from becoming a lucrative field for the medical profession. The whole medical armoury was tried at one time or another: heat, cold, bleeding, cupping, purgatives, and more exotic personal remedies such as boiled horse dung and oil extracted from human bones. There were also many secret remedies. Some of these included colchicum (often called autumn crocus), which is effective in relieving gout. George IV demanded it in place of the violent remedies suggested by his physicians.

Although this is a very cheerful book — 'ludic', to use a favourite word of the authors — it also has some of the problems common to most works that try to hammer out signs and signifiers over the centuries through the examination of stray references. This is most apparent in the chapters on the nineteenth century, where the literary manifestations are considerably less interesting than those of the eighteenth. Gout still appeared as a rich man's malady and hence a convenient shorthand for the dissipated habits of the

ruling classes, but writers who endured the condition were much more reticent, leading Porter and Rousseau into much discussion of "suppressed" gout. The gouty Wilkie Collins afflicted the enervated squires in his novels with a gout-like, but nameless malady. Joseph Conrad, another martyr to gout, never mentioned it in his fiction. Neither Jane Austen nor Thackeray afflicted their upper classes with gout. Eliot's baronets are also in prime condition, and lazy mentally rather than physically. Gout was losing its imaginative hold. But gout remained a refuge for the writer, and an excuse for retreat from the world. This was particularly obvious in Coleridge and Tennyson, but also in Darwin, who named gout as one of the many ailments that allowed him to avoid society and concentrate on his books.

Sometimes the line of argument on suppressed gout, fascinating though it is, fails to convince. Admirers of George Eliot's *Middlemarch* will not recognize Porter and Rousseau's imaginative interpretation. They will not think of the small town of Middlemarch as antithetical to the surrounding country, but of the "movement and mixture" between them that Eliot describes; they will expect to find Lydgate the nephew, not the son, of a baronet (indeed, in *Gout* he appears as both son and nephew in the same paragraph); they will be irritated by the references to "Ros", since Rosamund was never so abbreviated by her creator; they will think that the solitary comment on gout as a disease with "a good deal of wealth on its side" appears at the end of the book, not at the beginning; they will believe that Mr Casaubon was called Edward, not Isaac; that Lydgate does not retreat from medical practice to write on gout, but uses the reputation he has gained from his treatise on the subject to attract wealthy patients; and that he does not die of a possible "suppression" of gout, but of diphtheria. Perhaps the authors have not been using the standard 1874 edition, but it will be hard to convince Eliot's admirers that gout is a culminating symbol of this mighty work.

Scientific sources for the nineteenth-century, however, are more plentiful, and the scientific discussion is straightforward and absorbing. In mid-century, Alfred Baring Garrod showed how to detect uric acid in the blood, and linked it clearly with gout. Garrod was also firm on the role of alcohol. But, as the authors show, his work did not stifle the debate on how far hereditary factors were involved, nor on whether gout was treatable.

The last two chapters are not chronological but discursive, analysing the language of gout in terms of contemporary political and social perceptions and through the visual heritage. The excellent illustrations are concentrated on the eighteenth century, and reflect the comic, aristocratic and self-indulgent themes of the narrative. There is only

one twentieth-century cartoon, more meaningful in the American context, again suggesting that gout is a financial boon to the medical profession.

Susan Sontag argued, at the height of the AIDS panic, that illness should be seen as a scientific category, not a cultural or moral stigma. In their brief epilogue, Porter and Rousseau argue the contrary, that social metaphors may enable the patient to cope with disease. The pain of gout was bearable because of its excellent pedigree and its apparent promise of insulation from worse maladies. Yet the book also shows how regularly, and erroneously, medical men believed that they had understood this elusive complaint. Underneath its fashionable phraseology, which the reader will appreciate according to taste, this entertaining book succeeds very well as an old-fashioned treatise on medical hubris. □

Anne Crowther is in the Department of Economic and Social History, University of Glasgow, Glasgow G12 8QQ, UK.

Liberator or 'fix'?

On the Pill: A Social History of Oral Contraceptives 1950–1970

by Elizabeth Siegel Watkins

Johns Hopkins University Press: 1998. 202 pp. \$25.95, £21.50 (pbk)

Richard Davenport-Hines

In 1960, the US Food and Drug Administration accepted the G. D. Searle pharmaceutical company's Enovid pill as an oral contraceptive. By the thirtieth anniversary of this event, 80 per cent of US women born since 1945 had used contraceptive pills at some time in their lives. The pill had been swallowed as a daily routine by more humans than any other prescribed medicine. The US *Ladies' Home Journal* declared on this anniversary that the pill had "transformed our lives" even more than "winning the right to vote", while *The Economist* included the pill as one of the seven wonders of the modern world.

Elizabeth Siegel Watkins set herself the tasks of testing these claims in her intelligent and well-structured monograph, and of documenting changing perceptions of oral contraception from a "private vice" to a "public virtue" and finally to an issue of individual discretion. There is one drawback to her approach. Her evaluation is exclusively concerned, by her own admission, with "the impact of the pill on middle-class American society". In confining her researches to the United States, and by focusing on the experiences and practices of that nation's more articulate citizens, her work can seem insular. Bernard Asbell's *The Pill: A Biography of the Drug That Changed the World* (Random House, 1995) is more journalistic, but his

careful placing of US experiences in an international context provides a more suggestive treatment of social policy issues than Watkins.

In 1951, the feminist American philanthropists Margaret Sanger and Katherine McCormick commissioned the scientist Gregory Pincus to develop an infallible oral contraceptive, intended to liberate women's sexual acts from anxieties about their fertility. This stimulated other researchers to enter the field. In 1957, the G. D. Searle pharmaceutical company began marketing Enovid, ostensibly to treat gynaecological disorders. Its anovulant and therefore contraceptive properties became so well known, however, that by 1959 half a million US women were using it.

Contrary to widespread public and professional beliefs, the contraceptive revolution of the 1960s did not cause a sexual revolution. As demographers analysed the contraceptive habits of married women to document the contraceptive revolution (the pill was unavailable to unmarried women in many areas until the early 1970s), sociologists surveyed the sexual practices of unmarried women to depict long-term changes in sexual behaviour, and journalists bastardized their findings to present their caricature of 'the swinging sixties'. Yet years before Enovid, Alfred Kinsey and other sexual researchers had reported rates of pre-marital sexual intercourse steadily rising since the late nineteenth century. During the 1950s, the US marriage rate reached an all-time high, and the average age at which people married reached its all-time low; by 1959 almost half of brides were aged 18 or less. Many did not want to be burdened immediately with children, but reliable contraceptive information and technology was often unavailable. American puritanism flourished then as now. In 1960, 30 states of the union retained statutes prohibiting or restricting the sale or advertisement of contraceptives. Only in 1972 did the Supreme Court declare unconstitutional a Massachusetts law prohibiting the sale of contraceptives to unmarried people.

Watkins summarizes medical controversies surrounding the pill's safety. From the 1960s, medical studies linked the pill with an increased risk of strokes and breast cancer (although other studies reported that oral contraceptives protected against uterine and ovarian cancers). The validity of these findings has never been conclusively settled. Indeed, during the 1970s, the subject was confused by an often sensationalist 'media blitz', and mired by the intervention of individuals who objected to 'planned parenthood' on religious grounds, or to sexually independent women for other reasons. As a result of these 'health scares', by 1988 almost half of US married couples relied on either male or female sterilization to avoid preg-

nancy. American physicians' fear of malpractice suits raised important issues of 'informed consent' when prescribing oral contraceptives.

Some American feminists also turned against the pill in the 1970s. They regarded it as a technological 'fix' which did not address fundamental issues of oppression. Resenting any form of birth control kept within the jurisdiction of the medical profession, they advocated the diaphragm and cervical cap as barrier contraceptives that women could personally administer. The identification in the 1980s of the sexual transmission of HIV led to a revival in the popularity of condoms. In the early 1990s, Norplant, a subdermal implant that releases a synthetic hormone into the blood, and Depo-Provera, a hormone injection with contraceptive effect, have become available in the United States. However, further contraceptive innovations are unlikely to be developed in the United States. All but one of its pharmaceutical companies were scared out of contraceptive research and development in the 1980s by the intolerable litigiousness of American society.

On the Pill, which contains splendid illustrations, is, within its declared limits, an admirable exercise in social history. It depicts the cultural and ideological pressures on US medicine while demonstrating why about 19 million American women still use the pill in 1998. □

Richard Davenport-Hines is at 51 Elsham Road, London W14 8HD, UK.

Even a loose cannon may hit the right spot

Dancing Naked in the Mind Field

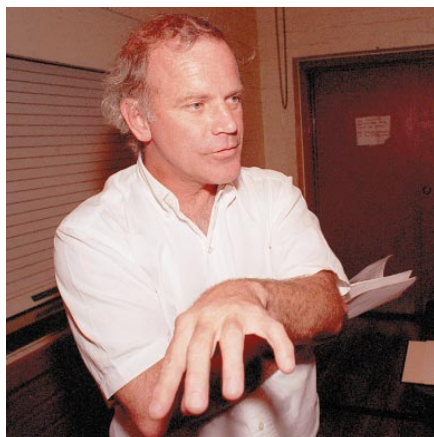
by Kary Mullis

Pantheon: 1998. 222 pp. \$24

Daniel S. Greenberg

In addition to scientific immortality and a wad of cash, the Nobel Prize provides an irrevocable licence to pontificate publicly on any topic, relevant or not to the recipient's expertise. Winners at other great competitions, for example, the Academy Awards and the Miss America contest, may assume an unrestricted right to mount the soap box and pronounce on issues of the day. But, for drawing a respectful, guaranteed audience, no honour can match the Nobel Prize. And no Nobel laureate comes close to Kary Mullis in the exercise of the accompanying pontifical rights.

Mullis earned a place in scientific history in 1983 as the inventor of the polymerase chain reaction (PCR), which quickly became the indispensable laboratory technique for genetics research. For this achievement, he shared the 1993 Nobel Prize for



Mullis: as a Nobel laureate “no door in the world will fail to open for you at least once”.

chemistry. Beyond that distinction, Mullis stands out for a number of reasons, including behaviour that oscillates between merely eccentric and obnoxious, and utterances that, when not loony, are reminiscent of the child who exclaimed on the emperor's state of undress. He also has an unrestrained penchant for faecal and copulatory terms.

Without the Nobel Prize, proclaimed on a book jacket that pictures Mullis shirtless with surfboard, it is doubtful whether this bizarre pottage would ever have found a publisher. But here it is, an easy weekend read that includes charming recollections of his childhood tinkering with chemistry, hair-raising accounts of trips on LSD and other drugs, disbelief in HIV as the cause of AIDS, and denunciations of those who do not share his respect for astrology. (“There's no proven body of facts in the social sciences”, he asserts, “that says human behavior does not contain elements that are related to planetary positions at the time of birth.”)

Along the way, Mullis also provides pithy insights into the workings of modern science, observing, for instance: “Probably the most important scientific development of the twentieth century is that economics replaced curiosity as the driving force behind research.” And: “When the National Institutes of Health makes an announcement through one of its many spokespeople, who checks out the credibility of that statement? Checks and balances are hard to come by in a scientific establishment that is supported from outside by a populace unskilled in the scientific arts.”

Mullis recounts the well-known tale of how the PCR breakthrough occurred to him during a long, night-time drive to his northern California cabin, a girlfriend at his side — one of many girlfriends sprinkled throughout his book, along with the three wives who preceded his current spouse.

Mullis acknowledges the Nobel Prize's power to suspend critical judgement in otherwise sensible people he encounters. “Once you have been given that accolade,” he

notes, “no door in the world will fail to open for you at least once. It is a free pass for the rest of your life” — even in the case of Mullis, self-described as “a loose cannon on the deck”.

His recognition of the “at least once” limitation on doors opening for Nobel laureates is based on experience. Several years ago, Mullis was invited to lecture on PCR to the European Society for Clinical Investigation. According to an indignant report by the outraged president of that organization, Mullis's “only slides (or what he called his art) were photographs he had taken of naked women with colored lights projected upon their bodies”. The president added that, in remarks to the audience, Mullis “accused science of being universally corrupt with widespread falsification of data to obtain grants”. In a published warning to colleagues, the president declared that his society “will not be inviting Dr Mullis to further meetings”.

For those who might be similarly offended by his words and slides, Mullis later announced that for a minimum of \$500 he would refrain from lecturing at any institution. He explains that he derived the concept of payment for not appearing from his experience with the Glaxo pharmaceutical company, which had acquired Burroughs Wellcome, the manufacturer of AZT. Glaxo, he writes, had offered him a \$1,500 speaking fee in 1993. When he responded that it was not enough, Mullis reports, Glaxo accepted his demand for \$3,000 and two first-class air fares. Glaxo then cancelled the invitation, he writes, when it learned that “I would speak about the fact that there is no scientific evidence that HIV is the probable cause of AIDS and that I believed people taking the drug AZT were being poisoned.” Whereupon, Mullis continues, he demanded \$6,048 to compensate for loss of “income from other potential engagements” that he had coupled to the lecture trip. Glaxo promptly paid that amount, he reports, providing the inspiration for a Mullis programme entitled “Have Slides, Will Stay Home. Yes ... But You Must Act Now ... Special Offer”.

Abandoning research, Mullis now writes, consults and lectures about science. The immense wealth generated by PCR eluded him. The Cetus Corporation, where he worked when he invented the PCR technique, paid him only \$10,000 for the patent, which it later sold to Hoffmann-La Roche for \$300 million. Neither firm has ever sent him a birthday card, he complains, adding, with characteristic Mullis bravado: “Screw Cetus and the Swiss”.

What might have happened if Mullis had captured the riches of PCR? Would he today be dispensing scores of philanthropic millions through the Kary Mullis Foundation? Who would get the money, and for what? Interesting to contemplate; or maybe horrible. □

Daniel S. Greenberg is at 3736 Kanawha St. NW, Washington, DC 20015, USA.

Seeking certainty in an unreliable world

The Golem at Large: What You Should Know about Technology

by Harry Collins and Trevor Pinch
Cambridge University Press: 1998. 155 pp.
£12.95, \$19.95

Barry Barnes

Of the many impressive texts that use case studies to convey ‘what you should know about technology’, *The Golem at Large* is the clearest and simplest. The authors rework existing materials with great care to produce a valuable introduction to their topic that is accessible to anyone. It is, however, necessary to clarify just what that topic is. The case studies presented here are all controversies, about the efficacy of technological artefacts, or the adequacy of technical knowledge or advice. Harry Collins and Trevor Pinch are concerned with the unreliability of what is generally regarded as, and indeed found to be, reliable. They do not focus that concern merely on technology: only three of the seven case studies relate to the working of artefacts; the others describe debates between scientists or technical experts.

The message of the book is that experts are fallible and liable to make mistakes. The claim may seem unremarkable, and is often stressed by experts themselves. But Collins and Pinch insist that a widespread image of scientific knowledge as certain, and technological devices as unconditionally trustworthy, needs to be opposed. Perhaps they are right: memories linger of how BSE, or ‘mad cow disease’, was said to pose “no conceivable risk” to humans, and many similarly ill-judged remarks are quoted here. On the other hand, perhaps they exaggerate the importance of this myth of the certainty of science as a foil for their own arguments. Either way, it is worth asking whether a book designed to attack claims of certainty and omniscience would not be better entitled



Crash test: technological devices should not be trusted unconditionally, say Collins and Pinch.

'what you should know about propaganda'.

Collins and Pinch are relativist sociologists who now seem to be widely regarded as among the most radical critics of natural science and scientific expertise, or at least have been denounced as such in the context of the 'science wars'. At first sight, this book confirms the diagnosis: it is 'biased' towards unedifying controversy and technical failure; it includes lengthy discussions of major disasters, such as Chernobyl and the explosion of the US space shuttle; it presents cases in which lay knowledge proves superior to that of experts. But once attention shifts to how the discussion actually proceeds, a very different story suggests itself.

To read these studies is to read what in the last analysis is a powerful defence of experts and expertise. Indeed, I know of no book more sympathetic to them outside the domain of hagiography. The authors stress the formidable difficulty of applying expertise, and the inordinate complexity of the real-world situations in which it is applied, and thereby seek to expose the hindsight-based critical attacks on experts, invariably occasioned when 'things go wrong', as facile, ill-informed and frequently self-serving. It is intriguing to find Collins and Pinch cheerfully using available 'best knowledge' as the basis of their own accounts of 'what really happened'; and even lamenting the absence, in one of the situations they describe, of "compelling evidence" of the kind available in astronomy.

There is nothing significantly critical of science, technology or expertise in this book. Indeed, its approach is profoundly conservative. Expertise is going to go wrong, but that is the nature of the beast. A touch of additional reserve and scepticism may be in order, a certain reflectiveness perhaps in the face of expert pronouncements, but nothing else: "We offer no policies". Even the criticism of the myth of scientific certainty is offered only for the greater good of science itself: if we expect too much of science and technology there is the danger of disillusion, of a "flight from reason", a "fall back into a dark age". In the light of all this, it is tempting to suggest that scientists do not always recognize who their friends are, although in the science wars, of course, it might have been that some scientists felt a need for enemies, and that the likes of Collins and Pinch were all they could manage to find.

It is likely that this book, like the authors' similarly designed collection of scientific controversies published five years ago, will be taken up for teaching purposes at an elementary level. The studies should stimulate valuable reflection in this kind of context, and may be used to illustrate far more themes than those discussed explicitly herein. None the less, the very narrow focus of their own discussion might be thought a disadvantage.

It is conceivable that the book will be

'balanced' in some contexts with materials stressing the positive achievements and exemplary reliability of science and technology, and 'balanced' in others with more forthright challenges to their authority and value; certainly, current trends in the mainstream of sociology are in the latter direction. Introductory textbooks, long notorious for making scarcely any mention of science and technology, are now beginning to refer to them, but mainly negatively, in relation to the rise of an alleged risk society and an increasingly threatened environment. But why are Collins and Pinch in practice so very much more positive about technical expertise than is now normal in their discipline? Perhaps part of the answer lies in the many years of close contact they have had with scientists and experts in the course of their substantive research. □

Barry Barnes is in the Department of Sociology, Exeter University, Rennes Drive, Exeter EX4 4RJ, UK.

Causing a buzz

Honeybees of Africa

by Howard R. Hepburn and S. E. Radloff
Springer: 1998. 370 pp. £64.50, \$109

Thomas D. Seeley

The honeybee, *Apis mellifera*, is the familiar species used for most of the world's beekeeping. It is native to Europe, the Middle East and the whole of Africa, and has been introduced by beekeepers to the Americas, Asia, Australia and the Pacific islands. Although most scientific studies of the honeybee have been conducted in Europe and North America, and have involved the subspecies native to Europe, more than two-thirds of this species' natural distribution area falls in Africa. Across this continent the honeybee inhabits such ecologically diverse settings as lowland rainforests, semi-arid savannahs, steamy coastal swamps and cool mountain ranges.

The smallest and the largest, the blackest and the brightest, and the gentlest and the

fiercest forms of *A. mellifera* exist in Africa.

The authors of *Honeybees of Africa* describe the honeybee scene in Africa as "a magnificent natural experiment", offering a special opportunity to investigate the nature of gene flow, population structure and biological adaptation. This is because these bees have been disturbed by humans only as honey-hunters and fire-starters; more disruptive interventions such as migratory beekeeping and selective breeding are virtually unknown in most of the continent.

The book begins with one of its most significant contributions: a detailed re-examination of the subspecies classification of the honeybees of Africa based on a new multivariate morphometric analysis. By amalgamating their own database at Rhodes University in South Africa with that of the late Friederich Ruttner of the Institut für Bienenkunde in Germany, the authors were able to base their analysis on 18,175 worker bees, representing 1,000 colonies in 291 localities across the continent. The result is a superb presentation, region by region (Maghreb, Nile Valley, East Africa and so on), of the geographical variability and population structure of the African honeybees.

This morphometric analysis, together with genetic studies by others, shows that populations of honeybees thought to be homogeneous and thus defined as subspecies actually show complex geographical variation. A recurrent theme is, therefore, the problem of accommodating natural population variation in a classification scheme. Unfortunately, to name things (such as populations of honeybees) typifies them and leads to typological thinking. In the end, the authors essentially follow Ruttner's classification system of subspecies names, because names are needed to discuss things, but the reader is shown clearly the tremendous variation within the honeybee populations of the continent.

The book's other major contribution is a comprehensive review of the scientific literature. For many topics — including the seasonal migration of colonies, the curious fertility of workers in queenless colonies of *A. m. capensis*, and the diverse predators and parasites of honeybees — the authors provide the best summary available. Often, though, the discussion is necessarily thin, simply for lack of information.

Honeybees of Africa provides biologists with an excellent source of information and challenges. I applaud the authors for thoroughly synthesizing what is known about *A. mellifera* across the whole of Africa, thereby setting the stage for countless exciting discoveries. □

Thomas D. Seeley is in the Section of Neurobiology and Behavior, Cornell University, Mudd Hall, Ithaca, New York 14853, USA.



Experimental honeypot: African honeybees can teach us a lot about populations and gene flow.

Metapopulation dynamics

Ilkka Hanski

Metapopulation biology is concerned with the dynamic consequences of migration among local populations and the conditions of regional persistence of species with unstable local populations. Well established effects of habitat patch area and isolation on migration, colonization and population extinction have now become integrated with classic metapopulation dynamics. This has led to models that can be used to predict the movement patterns of individuals, the dynamics of species, and the distributional patterns in multispecies communities in real fragmented landscapes.

Spatial structure of populations was a key element in some early concepts and models of population ecology^{1,2}, genetics^{3,4} and adaptive evolution⁴. In the past decade, the implications of spatial structure and dynamics have become widely recognized across population biology. The attention of ecologists has now been captivated by models^{5–7} demonstrating the profound influence of spatial locations of individuals, populations and communities on their dynamics: the essence of spatial ecology is that the spatial structure of ecological interactions affects populations as much as do average birth and death rates, competition and predation. The rapid destruction of natural habitats has highlighted the importance of spatially explicit ecological models.

I distinguish between three approaches to large-scale spatial ecology (Fig. 1). Theoretical ecologists have investigated a range of models depicting individuals with localized interactions and restricted movement range in uniform space^{6–8}, demonstrating how population dynamic processes can generate complex dynamics and spatial patterns without any environmental heterogeneity. By contrast, landscape ecologists have been occupied by descriptions of the generally very complex physical structure of real environments and the movement of individuals and resources in them^{9,10}. Symptomatically, the line of research pursued by theoretical ecologists is short of testable model predictions, whereas landscape ecology lacks a convincing theoretical framework. The third approach, metapopulation ecology, attempts to strike a compromise: here landscapes are viewed as networks of idealized habitat patches (fragments) in which species occur as discrete local populations connected by migration. Many species live in such well delimited habitat patches (ponds, woodlands in agricultural landscapes, and so on) that the 'patch network' assumption (Fig. 1) is really no simplification at all; for other species, it is a useful approximation; whereas for others, it is unhelpful because these species have a more continuous popula-

tion structure in space.

The kind of metapopulation approach depicted in Fig. 1 does not include everything that is commonly assigned to metapopulation ecology. Other approaches have been reviewed, for instance theoretical work based on lattice models^{6,7,11–13} and conservation-oriented empirical studies^{14–16}. The term 'metapopulation' is often used for any spatially structured population, and 'metapopulation dynamics' then covers all spatial dynamics. My agenda here is more restricted: I aim to highlight the ways in which the patch network assumption (Fig. 1) has facilitated the concurrent development of models and empirical research and so provided insight into the dynamics of real metapopulations in highly fragmented landscapes.

The foundation of the current metapopulation concept is in Levins's¹⁷ vision of a metapopulation as a 'population' of unstable local populations, inhabiting discrete habitat patches such as those shown in Fig. 1. A classic metapopulation persists, like an ordinary population of mortal individuals, in a balance between 'deaths' (local extinctions) and 'births' (establishment of new populations at unoccupied sites). In this respect, metapopulation ecology shares similar conceptual underpinnings with epidemiology^{18,19}: susceptible and infected individuals represent empty and occupied 'patches' for parasites. Some key results are essentially the same, for instance the critical community size for stochastic persistence of infectious diseases²⁰ can be compared with both patch-area-dependent extinction of local populations and patch-number-dependent extinction of metapopulations of free-living organisms, as discussed below.

Metapopulation dynamics in a broad sense are not restricted to systems with population turnover, extinctions and colonizations, but the concept developed here is based on Levins's classic metapopulation idea with extinction-prone populations in discrete habitat patches. I first discuss the new perspective that metapopula-

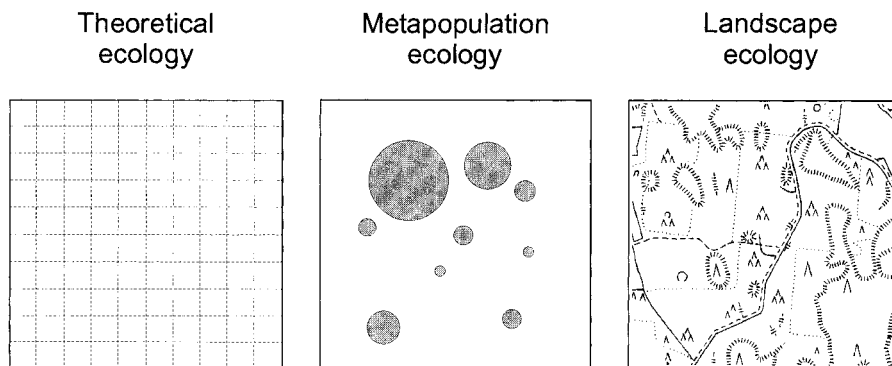


Figure 1 Three approaches to spatial ecology. Theoretical ecologists typically assume homogeneous continuous or discrete (lattice) space. Landscape ecologists tend to analyse the structure of complex real landscapes, with less emphasis on modelling population dynamics. Metapopulation ecology, in the

middle, makes the simplifying assumption that suitable habitat for the focal species occurs as a network of idealized habitat patches, varying in area, degree of isolation and quality (the latter is not shown or discussed here, but see ref. 77), and submerged in the midst of uniformly unsuitable habitat.

Table 1 Processes influencing extinction in metapopulations

Scale of extinction	Scale of process	Extinction due to stochasticity	Extinction due to extrinsic causes
Local extinction	Local processes	Demographic* Environmental Genetic*	Habitat loss Generalist enemies and competitors Persecution by humans etc.
Metapopulation extinction	Metapopulation processes	Migration in small populations Extinction-colonization Regional	Specialist enemies and competitors Habitat loss and fragmentation, extinction typically delayed

The processes that operate in the well studied Glanville fritillary butterfly metapopulation²¹ are printed in bold.

* Demographic and genetic stochasticity assume an increased significance in metapopulations with many small local populations.

tion ecology provides for population extinction, previously considered largely in the context of isolated populations. The general implications of extinction-colonization dynamics are then outlined. The basic model is augmented by assuming the kind of spatially realistic network structure shown in the middle panel in Fig. 1, with spatial variation in patch areas and degrees of isolation. The subsequent discussion is focused around the key question for ecology and conservation, the conditions under which metapopulations persist when habitat area is lost and the remaining habitat becomes ever more fragmented.

Extinction in metapopulations

In metapopulations, population extinction is a recurrent rather than a unique event, which adds to the range of extinction processes that have significance in nature and forces us to construct an increasingly mechanistic and biologically enriched view of extinctions. Table 1 summarizes the processes of extinction operating in metapopulations, some of which will be discussed below. Most of these processes have been documented in a large metapopulation of a well studied species, the Glanville fritillary butterfly, *Melitaea cinxia*²¹, underscoring the point that many processes typically contribute to extinctions in metapopulations.

The two familiar forms of stochasticity affecting local populations, demographic and environmental stochasticity, have exact counterparts at the metapopulation level in extinction-colonization and regional stochasticities²². To appreciate the significance of extinction-colonization stochasticity, let us first attend to the following necessary condition for long-term metapopulation persistence: the expected number of new populations generated by one existing population during its lifetime in an otherwise empty patch network must be greater than one²³. An analogous replacement condition naturally applies to individuals in local populations and is well developed in the epidemiological theory¹⁸. The replacement condition is necessary but not sufficient for long-term persistence. In a small metapopulation, all local populations may happen to go extinct at the same time owing to extinction-colonization stochasticity, even if the replacement condition is met, just as all individuals in a small population whose finite rate of population increase is greater than unity may happen to die without leaving any surviving progeny. In real patch networks, there is additionally the complication that not all patches are equally connected, hence whether the replacement condition is met or not may depend on the focal patch²⁴.

Studies on the Glanville fritillary butterfly have produced results that illustrate the operation of extinction-colonization stochasticity in metapopulations. In a comprehensive survey of 127 relatively independent patch networks, only a third of networks with less than 15 patches were occupied, whereas practically all larger networks were occupied²⁵, in agreement with theoretical predictions about extinction-colonization stochasticity²⁶. Not all absences from small patch networks are necessarily due to extinction-colonization stochasticity only, because some of the metapopulations may not satisfy the replacement condition and are hence expected to go extinct any way. Nonetheless, an important conclusion is that, whether the replacement condition is met or not, a metapopulation

of extinction-prone local populations in a small patch network is necessarily more threatened than are metapopulations in large and well connected networks.

Regional stochasticity²² influences the dynamics of many populations simultaneously and leads to spatially correlated population dynamics. Large-scale spatial synchrony may also be generated by other processes such as migration and predation⁷, but regional stochasticity in the form of spatially correlated weather conditions is probably the dominant synchronizing mechanism. Regional stochasticity reduces the number of effectively independent local populations²⁷ and, if strong enough, can make metapopulation-level persistence of classic metapopulations less likely.

Apart from the local processes, local extinction in metapopulations is influenced by processes that are best addressed at the metapopulation level. Emigration may reduce population growth rate and, in combination with demographic and environmental stochasticity, lead to increased risk of extinction. This is particularly likely to happen in small habitat patches with a frequently increased per capita emigration rate (ref. 28, and I.H., J. Alho and A. Moilanen, manuscript in preparation). Conversely, immigration from nearby large populations may reduce extinction risk in small populations²¹.

Metapopulation structure gives extra scope for demographic and genetic stochasticities to operate, because classic metapopulations

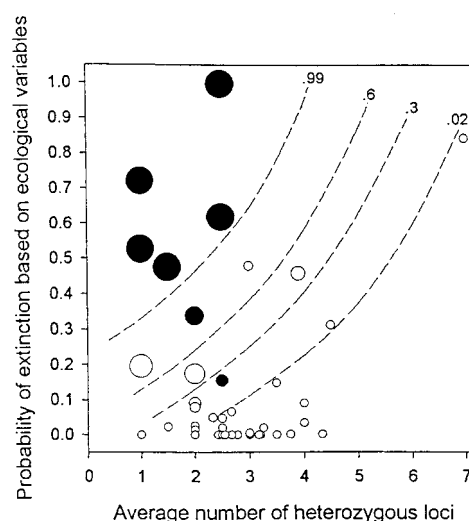


Figure 2 The probability of extinction in the Glanville fritillary butterfly is influenced both by ecological factors and by heterozygosity, which is here used as a measure of the level of inbreeding²⁹. The vertical axis gives the probability of extinction for 42 populations as predicted by a model including several ecological factors²⁹. The horizontal axis gives the average number of heterozygous loci per individual in a sample of eight polymorphic enzyme and microsatellite loci. The size of the symbol is proportional to the probability of extinction predicted by a model including both the ecological factors and heterozygosity (the isoclines of equal extinction risk were drawn by eye). Of the 42 populations studied, seven populations (black) went extinct in one year²⁹.

typically include many small populations in which these forms of stochasticity may have a great impact. A recent study²⁹ on the Glanville fritillary butterfly demonstrated a significant increase in extinction risk with increasing level of inbreeding (Fig. 2). This finding appears to be contrary to the common wisdom that populations with a history of bottlenecks should not suffer from inbreeding, because such populations have supposedly become purged of deleterious recessive alleles, or have gone extinct³⁰. The Glanville fritillary metapopulation has not only gone through bottlenecks—the metapopulation literally exists in bottlenecks, because most individuals reside in very small local populations²¹. Theory and intuition developed to explain the effect of genetic stochasticity in isolated populations are inadequate for highly structured metapopulations. For instance, weakly deleterious alleles may accumulate in small populations because of genetic drift. In a metapopulation, different local populations will by chance carry different deleterious alleles, which become transferred between populations by migration. Present theory is inadequate to predict the ultimate outcome of these processes, and there is an urgent need to construct a more comprehensive theory of genetic stochasticity for highly fragmented metapopulations and to develop field projects to increase our empirical knowledge.

One of the greatest virtues of the metapopulation approach to population extinction is the opportunity to contribute to a mechanistic understanding of the biological consequences of habitat loss and fragmentation, the greatest threat to biodiversity worldwide. This will be addressed below, following a brief review of the general implications of extinction–colonization dynamics.

Extinction–colonization dynamics

Metapopulations consisting of small extinction-prone local populations can only persist regionally, in a balance between local extinctions and colonizations^{17,25}. Long-term persistence of classic metapopulations is essentially due to asynchrony in the dynamics of local populations, which reduces the variance in the intrinsic rate of metapopulation increase and hence reduces the risk of metapopulation extinction (applying theory developed for single populations^{31,32}). This type of persistence is very different from the one traditionally studied by population ecologists using life table and key factor analyses, where the ecologist's task is to uncover the density-dependent processes that lead to the regulation of individual populations^{33,34}. Nonetheless, the conclusion should not be drawn that metapopulation dynamics allow long-term persistence without any local density dependence, as has been repeatedly but mistakenly assumed for the past 40 years³⁵. Some density dependence operating in local populations is necessary for long-term persistence, even if metapopulation persistence is compatible with infrequent local density dependence in species with ephemeral local populations²¹.

Extensive theoretical literature has explored the population dynamic, genetic and evolutionary consequences of extinction–colonization dynamics^{6–8,21,25,36–41}. For ecology and conservation, a key consideration is whether the expected growth rate of a metapopulation in an almost empty network is positive, in other words whether the replacement condition referred to above is satisfied. The expected growth rate of a small metapopulation depends both on the intrinsic attributes of the species (local dynamics and migration behaviour) and on the spatial configuration of the habitat. For comparative purposes, fragmented landscapes can be characterized by measures of 'colonization potential' (Box 1), but a quantitatively valid estimate of metapopulation growth rate requires a spatially realistic metapopulation model.

The frequency with which species persist as metapopulations in extinction–colonization balance, as opposed to persisting as a result of stable large populations, is still debated among ecologists^{14,25}. Given that a large fraction of species on Earth are highly specific in their habitat requirements and are generally uncommon, it is

reasonable to guess that metapopulation dynamics represent a valuable and even necessary approach to a satisfactory understanding of the dynamics of a large number of species in many regions^{21,42}. However, there are so few comprehensive studies on particular species that it remains possible to argue both ways. For parasites

Box 1 Calculations

Parameter estimation and simulation of the incidence function model (IFM).

In the examples in Fig. 3, the IFM was parameterized with one snap-shot of patch occupancy data, as displayed in Fig. 3a, using maximum likelihood regression, assuming independent patch dynamics (Fig. 3b), and making an informed guess about the value of the minimum patch area for which the probability of extinction in unit time equals one^{48,49}. Parameter estimation for the three species is discussed in depth elsewhere (the Glanville fritillary^{48,50,69}, the American pika⁵¹, and the European nuthatch⁴⁹). For parameter estimation based on a Monte Carlo method and incorporating all spatial and temporal autocorrelations in patch occupancy, see ref. 50.

Having estimated the model parameters, one may numerically simulate extinction–colonization dynamics in any patch network, by assigning area-dependent extinctions and isolation-dependent colonizations independently for each patch in each generation using the probabilities given by the model^{21,48}. Regional stochasticity (spatially correlated environmental stochasticity) can be included by multiplying all patch areas in each time step with the same lognormally distributed random variable with mean 1 and standard deviation σ , which will translate into correlated temporal changes in extinction and colonization rates in metapopulation dynamics^{26,51}.

Deterministic equilibrium of the IFM. Apart from calculating the incidences (long-term probabilities of patch occupancy) by simulation, one may calculate the respective deterministic values by iterating the equation

$$S_i = \sum_{j \neq i} \exp(-\alpha d_{ij}) p_j A_j$$

$$\text{where } p_j = \frac{C_j}{C_j + (1 - C_j) \frac{e}{A_j^2}} \text{ and } C_i = \frac{S_i^2}{S_i^2 + y^2}$$

until convergence^{21,66}. p_j is the probability of patch j being occupied, d_{ij} is the distance between patches i and j , A_j is the area of patch j , S_i and C_i are the connectivity (Box 2) and the probability of colonization of patch i , respectively, and α , e , x and y are model parameters. To check for alternative equilibria, the iteration was started with either all p values equal to 1, or all except one equal to 0 (repeated for each i and averaged to obtain the initial values of S_i).

Colonization potential of a patch network. Whether a species is able to persist in a patch network or not depends on the amount of habitat in and the spatial configuration of the network as well as on the attributes of the species. For comparative purposes, networks can be characterized by their 'colonization potential', which is here calculated as $M = \bar{S}$, where S is the measure of connectivity in Box 2, now calculated on the assumption that all patches are occupied (because M reflects the potential, not the realized, rate of colonization). Note that of the attributes of the species, M includes only the migration range ($1/\alpha$), and is hence only an approximation. In the application in Fig. 4, the empirically observed scaling $N_j = A_j^{0.5}$ was used²⁶, and M was square-root transformed to spread out the data points more evenly along the horizontal axis.

Calculating the time delay in metapopulation response to habitat destruction.

The following calculation gives a practical measure of the length of the delay. Calculate first the deterministic incidences before habitat destruction, say J_1 , as explained above. Next calculate the expected number of occupied patches following habitat destruction, say P_2 , as the sum of the respective deterministic incidences. Then start a simulation in the reduced network but initializing patch occupancy with probabilities J_1 , and calculate the time until the number of occupied patches drops at or below P_2 . Repeat the simulation 100 times and use the median value as a measure of the delay.

living in 'networks' of host individuals, extinction–colonization dynamics are self-evident owing to the limited lifespan of host individuals.

Although this review is concerned with ecological dynamics, it is appropriate to mention that metapopulation dynamics have profound implications for the genetic structure and dynamics of species, potentially also for the adaptive evolution of species, even if Sewall Wright's⁴ 'shifting balance process' is no longer considered to be of wide importance in evolution^{39,43}. A key genetic question is the degree to which extinction–colonization dynamics decrease the effective genetic population size^{39,40}. Convincing empirical evidence is lacking on whether there is reduced genetic variability in species with a classic metapopulation structure.

If the rates of population extinction or colonization are influenced by genetically determined traits of populations, natural selection may operate at the level of local populations (group selection), either reinforcing or opposing selection at the level of individuals. As an example, the Glanville fritillary butterfly has two host plants in Finland, with a geographic variation in genetically determined oviposition preference (host plant choice), corresponding to a geographic trend in host plant abundances (M. Kuussaari, M. Singer and I.H., manuscript in preparation). Recent results indicate that the oviposition preference influences the probability of colonization of habitat patches with different mixtures of the two host plants (I.H. and M. Singer, unpublished results), such that butterflies from *Veronica*-preferring populations have a smaller probability of colonizing a *Plantago*-dominated habitat patch than have butterflies originating from *Plantago*-preferring populations. A model shows that, depending on the spatial structure and plant species composition of a patch network, group selection (biased colonizations) may either increase or decrease preference for a particular host plant in a metapopulation. The model-predicted shift in oviposition preference in particular metapopulations explained a large amount of the observed variation in host plant use, suggesting that group selection has influenced the oviposition preference in the Glanville fritillary metapopulations (I.H. and M. Singer, unpublished results).

Of the life history traits that are thought to be influenced by metapopulation-level selection⁴¹, the evolution of migration rate has a special significance because it directly influences metapopulation dynamics. Evolutionary ecologists have investigated whether species are likely to evolve an elevated migration rate in response to habitat destruction⁴¹, which might help species survive in increasingly fragmented landscapes. This may happen⁴⁴, but, given the current rate of environmental change, it seems more likely that species will go extinct before significant evolutionary changes have had time to occur.

The patch area and isolation paradigm

Real metapopulations do not consist of identical and equally connected populations as is assumed in the basic models¹⁴. Recognizing differences among local populations is important, but this does not mean that the original concept of metapopulation as a 'population of populations'¹⁷ would become obsolete. On the contrary, a new synthesis is taking shape in which the classic metapopulation concept is enriched with general effects of patch area and isolation on movements of individuals, extinction of populations and the establishment of new populations at empty sites (Box 2).

Individuals born to a metapopulation may stay all their life in the natal habitat patch or they may move one or more times to a new patch during their lifetime. To estimate the probabilities of survival and migration per unit time with mark-recapture data, a model of individuals' capture histories may be constructed based on the effects of patch area on migration and isolation on mortality during migration (I.H., J. Alho and A. Moilanen, manuscript in preparation). Because it is reasonable to assume that mortality within

habitat patches does not depend on isolation, unlike mortality during migration, one can tease apart, at least in principle, the two kinds of mortality. This is of interest to ecologists, because mortality during migration is an important cost of migration and has been hard to measure for most species. In a case study on the butterfly *Melitaea diamina*⁴⁵, the model was applied to a metapopulation consisting of 14 local populations within an area of 4 km², using data on 557 marked individuals with 1,301 recaptures. Immigration and emigration scaled as patch area to power 0.2, roughly half of the daily losses of individuals from large habitat patches (~1 ha) were due to emigration, only <1% of daily migration distances were

Box 2 Incorporating habitat patch areas and isolations into metapopulation models

Real fragmented landscapes typically show much spatial variation in patch areas and isolations. Individual and population processes and community patterns are generally influenced by patch area and isolation, and it is desirable to include these effects into metapopulation models, especially because such models can often be parameterized with data that are readily available from field studies (ref. 48, and I.H., J. Alho and A. Moilanen, manuscript in preparation). Adding the patch area and isolation effects into patch occupancy models of metapopulation dynamics has promoted a close link between modelling and field studies^{21,66}.

The scaling of individual and population processes and community patterns with patch area (A) has been described by the power function: process or pattern $\propto A^{\beta}$. At the individual level, the process is migration rate (I.H. *et al.*, in preparation), at the population level it is extinction risk⁴⁸, and at the community level the pattern is species number⁶⁷. There is no simple connection between the three power functions, but they can each be justified as an approximation based on biologically reasonable assumptions or model predictions (the latter in the case of species number). At the population level, the extinction risk scales asymptotically as a power function of the population ceiling and hence of patch size in a diffusion approximation to an extinction model^{31,32}.

Isolation of a habitat patch or a local population from existing populations influences the rate of immigration to a patch and hence the probability of colonization of an empty patch. A sensible index of connectivity (inverse of isolation) of patch i is $S_i = \sum_j \exp(-\alpha d_{ij}) N_j$. Thus connectivity of patch i increases with decreasing distances (d_{ij}) from, and sizes (N_j) of, existing other populations. The species-specific migration range is given by $1/\alpha$. The distance d_{ij} may be the Euclidian distance or some more complex measure taking into account the influence of landscape structure on migration⁶⁸. If knowledge on patch occupancy only is available, N_j values may be crudely estimated by a power function of patch areas, multiplied by one for occupied and by zero for empty patches. To measure connectivity for emigrants, values of N_j are replaced by a power function of patch areas for all patches, whether occupied or not, which assumes that probability of colonization of patch i is proportional to A^{β} .

The probability of an emigrant from patch i surviving migration and the probability of an empty patch i becoming recolonized are increasing functions of the respective S_i values. In both cases, a reasonable simple choice for the functional form is the sigmoid model (ref. 48, and I.H. *et al.*, in preparation)

$$x_i = \frac{1}{1 + \left(\frac{y}{S_i}\right)^2}$$

where x_i is probability of survival during migration from patch i or the probability of colonization of an empty patch i . y is a parameter.

Specific models of migration (I.H. *et al.*, in preparation), metapopulation dynamics⁴⁸ and community pattern⁶⁶ can be constructed with the above (or other comparable) assumptions about the patch area and isolation effects.

greater than 1 km, and 16% of all deaths were estimated to have occurred during migration (I.H., J. Alho and A. Moilanen, manuscript in preparation; for a general discussion of movement patterns in metapopulations, see refs 46, 47).

The incidence function model (IFM)^{21,48} is a prime example of the value of incorporating the patch area and isolation effects into a basic metapopulation model. The IFM is built on the well supported assumptions that extinction risk of local populations decreases with increasing habitat patch area (because extinction risk decreases with increasing expected population size, which increases with patch area²¹), and that the colonization probability is a function of patch isolation from existing local populations (Box 2). From these assumptions, a Markov chain model for state transitions in individual habitat patches leads to an expression for the average long-term probability of patch occupancy, called the incidence, as a function of the area and isolation of the patch⁴⁸. The IFM thus relates the pattern of habitat occupancy to the structure of the fragmented landscape as reflected in patch areas and isolations (Fig. 1, middle panel). The IFM has only a few parameters, which facilitates parameter estimation. Of particular interest is that the model allows one to estimate the parameters of the spatial processes, local extinction and recolonization, with spatial data on patch occupancy, although more robust parameter estimates can be obtained if data are also available on population turnover^{48–50}.

The IFM has been used to model the metapopulation dynamics of plants, insects, small mammals and birds²¹. Figure 3a–c gives three examples, which are used below in a discussion of the consequences of habitat destruction for metapopulation dynamics. The IFM has been tested by comparing the observed population turnover rate with the predicted one^{49,51,52} and by testing predicted patterns of patch occupancy in landscapes other than the one from which parameters were estimated²¹. In one case, parameter values estimated for a species of butterfly were used successfully to predict the distribution of a related species in another patch network⁴⁵. This example illustrates the potential predictive power of the IFM to explain the distribution of species in fragmented landscapes.

The IFM can be used to investigate the scaling of extinction risk with habitat patch area, and thereby with population size, using patch occupancy data. This scaling gives an interesting opportunity to relate the parameters of local extinction models^{31,32} to the parameters of metapopulation models²¹. The scaling constant was related to body size in small mammals and birds but not in insects, which implies that the strength of environmental stochasticity increases with decreasing body size in vertebrates²¹.

Complex spatial dynamics

Theoretical studies have demonstrated that spatial population dynamics may generate complex spatial patterns in species abundances in the absence of any environmental heterogeneity^{6,7,38}. Complex spatial patterns in spatially explicit models are due to localized interactions and restricted migration range of individuals, localized interactions amplifying small-scale chance variation and short-range migration ‘memorizing’ the altered abundances⁵³. Complex spatial patterns occur in single-species models but are especially well documented for metapopulation models of inter-specific competition⁵³ and predator–prey dynamics^{12,16,38}. Unfortunately, because of formidable logistical problems and the ever-present spatial heterogeneity in real landscapes, the model predictions have remained largely untested (but see refs 54–56). The predictions stem from deterministic models, and the outcome of the deterministic processes interacting with environmental stochasticity remains to be determined.

Another type of complexity that may arise in metapopulation dynamics is alternative stable equilibria²³, of which one corresponds to metapopulation extinction and the other to a state with most of the habitat occupied (Fig. 4a). In finite patch networks, every metapopulation will ultimately go extinct because of extinction–

colonization stochasticity, but meanwhile the metapopulation may settle to a positive quasi-stable equilibrium (Box 1), which for practical purposes is stable in large and well connected networks. Alternative stable equilibria are generated by a positive feedback in metapopulation dynamics (the rescue effect), migration from existing large local populations increasing the sizes of, and thereby decreasing the risk of extinction in, nearby small populations⁵⁷. Propagule size-dependent success of colonization (local Allee effect) may also contribute to alternative equilibria in metapopulations.

An extensive data set on the Glanville fritillary butterfly from many semi-independent patch networks shows a pattern of habitat occupancy that is suggestive of alternative equilibria (Fig. 4b). These results can be modelled with the IFM parameterized for the butterfly (Fig. 3b) to predict the quasi-stable equilibria in particular patch networks (Box 1). The IFM includes the rescue effect and propagule size-dependent colonization success and may hence generate alternative equilibria. The modelling results suggest that 12 of the 66 patch networks have alternative quasi-stable equilibria (Fig. 4c). Simulation of metapopulation dynamics in the 66 networks yields snap-shots of habitat occupancy (Fig. 4d) that are similar to the observed one (Fig. 4b). In these simulations, a small probability of colonization from outside the network was assumed to prevent permanent metapopulation extinction. An especially noteworthy feature of the simulation results is that many networks with colonization potential just below the threshold for a positive equilibrium can nonetheless have a large fraction of the habitat occupied at a given point in time (Fig. 4d), a pattern that is also apparent in the empirical data (Fig. 4b). This result indicates that the dynamic force pushing the metapopulation towards extinction is weak in these networks, a situation with important consequences for the response of metapopulations to habitat destruction, as will be discussed next.

Response to habitat loss and fragmentation

Habitat destruction—involving downright loss of habitat, degrading habitat quality, and fragmentation of the remaining habitat—is by far the most significant cause of population and species extinction. For instance, of the endangered bird species in the world, habitat loss has been singled out as the main threat in 82% of the species⁵⁸. Predicting the consequences of habitat destruction is perhaps the greatest challenge for metapopulation dynamics.

Conservation-oriented ecologists have employed complex simulation models of multiple populations to assess the large-scale and long-term effects of habitat loss and fragmentation^{59,60}. Unfortunately, these models tend to have many untested assumptions and a large number of parameters that are difficult to estimate. Given that the practical task is generally to compare alternative scenarios of landscape change, rather than to predict quantitatively the extinction risk of a particular metapopulation, simpler and more robust models such as the IFM or state-transition models^{21,61} may be preferable to complex simulation models, especially for highly fragmented landscapes²¹. But for metapopulations including one or more very large populations, these patch occupancy models are inadequate because they ignore local dynamics, and hence some more complex approach is needed.

Studies of landscape connectivity and metapopulation dynamics employing simple strategic models have led to three general conclusions about metapopulation response to habitat destruction. First, landscape ecologists have suggested that the response is nonlinear, because habitat connectivity is lost in a highly nonlinear manner in simple scenarios of habitat destruction⁶². A metapopulation dynamic reason for a nonlinear response to habitat destruction is alternative equilibria—the positive equilibrium may be suddenly lost with increasing habitat loss (consider moving to the left in Fig. 4). Second, metapopulation decline in response to habitat destruction occurs with a time lag²⁶, dubbed the ‘extinction debt’³⁶. And

third, simple models suggest that the equilibrium amount of empty habitat in a landscape before habitat destruction equals the extinction threshold, the minimum amount of habitat required for long-term persistence⁶³, which has been termed the 'Levins rule'²⁶. Figure 3d–f examines these conclusions for metapopulations of an insect (the Glanville fritillary), a passerine bird (the European nuthatch) and a small mammal (the American pika), with the help of the IFM. To make examples as realistic as possible, the IFM was parameterized (Fig. 3b) with a snapshot of patch occupancy data (Fig. 3a) for each species. Three forms of habitat loss were used to generate hypothetical landscape scenarios: a contiguous region of the landscape may be lost; habitat fragments may be lost randomly; and individual fragments may lose area. In reality, habitat destruction is likely to involve a combination of these situations.

The Levins rule is applicable when habitat is lost by random elimination of patches. In the examples shown in Fig. 3d, the number of empty patches decreases with decreasing habitat availability in the nuthatch metapopulation, but in the butterfly and mammal metapopulations there is no systematic change, in agreement with the Levins rule, although there is substantial variance for different landscape configurations. When habitat is lost in large blocks, the number of empty patches tends to decrease with decreasing amount of habitat. In contrast, when each individual patch loses area, leading to the same number but smaller patches, the number of empty patches inevitably increases with a reduction in total habitat area (Fig. 3d), as such habitat loss will increase local extinction risk and reduce colonization rate. In conclusion, there is not much practical value in the Levins rule—metapopulation response is too specific to the particulars of habitat loss and to the

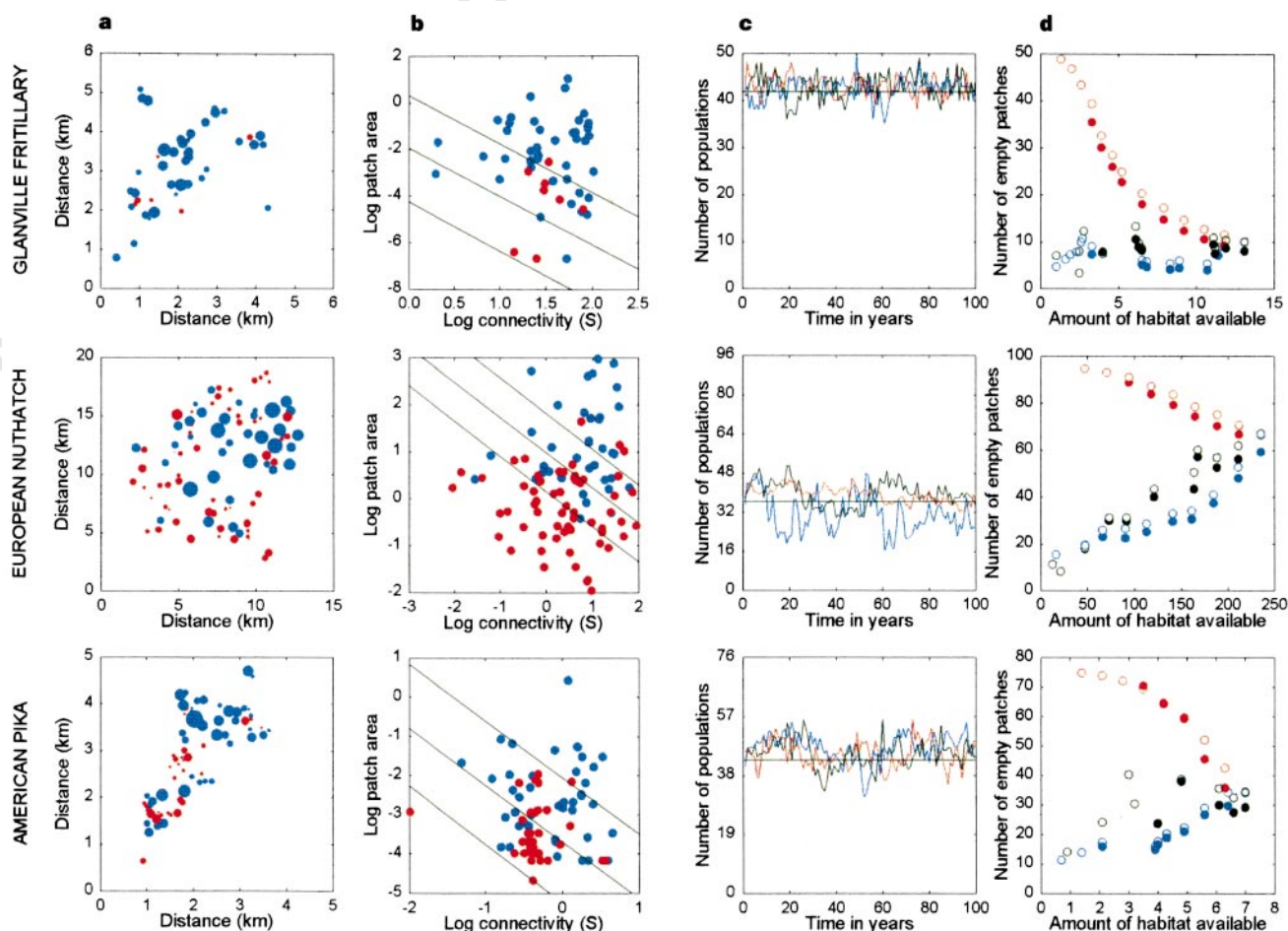
original fragmented landscape for the idealized model to be useful.

Metapopulations inevitably go extinct when the colonization potential of the landscape falls below a threshold, but stochasticity may lead to extinction even in landscapes with more and better connected habitat patches than specified by the threshold (Fig. 3e). In the small mammal example, the risk of extinction in 100 years increases rather gradually with decreasing amount of habitat, whereas in the butterfly and bird examples the response is more sudden. Note that in these examples the extinction risk is set chiefly by the amount of habitat remaining, not by the type of habitat destruction (Fig. 3e), which underscores the primary importance of the total area of suitable habitat.

Metapopulations track reduction in habitat area with a delay (Fig. 3f). The delay tends to become longer when the original amount of habitat (before the reduction) is smaller. The delay is especially long when the new equilibrium following habitat loss is metapopulation extinction (Fig. 3f). Long delays in these cases reflect the weak dynamic forces operating in landscapes that have colonization potential just below the threshold for persistence (Fig. 4d). The important message here is that many species may be hanging around for a long time in landscapes that have already lost their capacity to support these species on the long term.

Community patterns

Metapopulation models have been extended to interactive and non-interactive multispecies communities. Models of interacting metapopulations have been used to explain patterns in species succession⁶⁴, species richness and composition³⁶ and food web structure⁶⁵ of communities. Metapopulation dynamics may set a



limit to food-chain length in communities of specialist species in fragmented landscapes, as the density of suitable (prey-occupied) habitat patches is necessarily lower for high trophic levels⁶⁵. Predator–prey^{12,13,38} and competitive^{36,53,63} metapopulation dynamics will not be discussed here.

Community models for non-interactive species are constructed simply by summing up the predicted patterns for individual species. Using this approach, it has been demonstrated⁶⁶ that the two most general patterns in the distribution of species, the species–area (SA) curve⁶⁷ and the distribution–abundance (DA) relationship⁶⁸, which have been studied in complete isolation, can be derived from the same metapopulation model with area-dependent extinction rate and distance-dependent colonization rate (this demonstration by itself does not exclude alternative explanations).

Our model⁶⁶ was constructed for a community in which species have different densities, reflecting, for example, interspecific differences in the degree of ecological specialization. Other things being equal, higher density increases colonization rate by increasing the absolute number of immigrants, and decreases extinction rate by boosting population sizes. The metapopulation model leads to an expression for the incidence of species *i* on island (or habitat patch) *j*. Summing up the incidences across the species gives the expected number of species on islands, and regressing this number against island area gives the SA curve. Similarly, summing up the incidences across the islands gives the expected distribution of the species; and regressing the distribution against species' density leads to the DA curve. Figure 5 gives examples of model-predicted SA and DA curves and demonstrates how the slopes of these curves depend on a small number of parameters. The predicted slope values are testable

for landscapes and communities for which appropriate data can be collected.

Conservation and landscape management

Species have adapted to their environments and hence, according to the common wisdom in ecology, the distribution of species' abundances in space reflects the match between the environment and the species' ecological requirements, with an appropriate caveat for ecological interactions such as competition and predation and recent perturbations caused by humans. Spatial ecology challenges a strict interpretation of this habitat–organism relation, as species may exhibit complex spatial patterns in uniform environments; species may be absent where environmental conditions are favourable (owing to stochastic extinction of local populations); and species may be present where conditions are not favourable (sink populations). These general conclusions have profound implications for conservation, as do several more specific results stemming from research in metapopulation dynamics.

The total amount of habitat in the fragmented landscape is often a good predictor of long-term metapopulation persistence (Fig. 3e), though it is easy to construct counter-examples⁴⁵. Thus, in spite of all the interest in fragmented populations, the primary aim in conservation should be simply to preserve as much habitat as possible. Spatially realistic metapopulation models may be used^{45,51,59,60,69} to generate more refined species-specific and landscape-specific predictions, which replace the infamous rules of reserve design based on the dynamic theory of island biogeography⁷⁰. The emphasis in modelling should be in transient dynamics, to provide insight about the dynamics at the timescale at

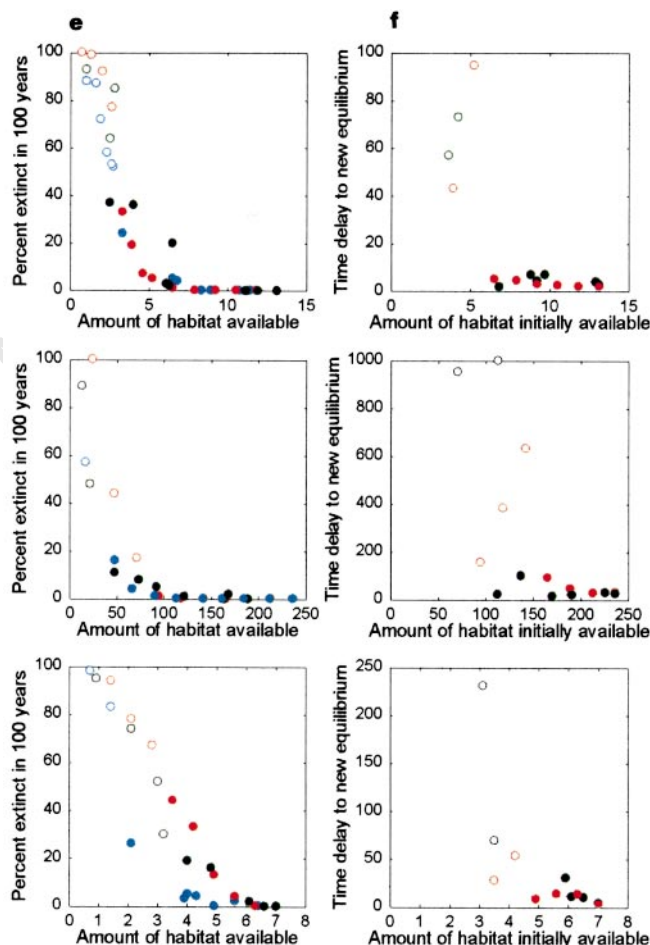


Figure 3 The incidence function model (IFM) is here parameterized for metapopulations of an insect (the Glanville fritillary butterfly, *Melitaea cinxia*), a passerine bird (the European nuthatch, *Sitta europaea*), and a small mammal (the American pika, *Ochotona princeps*), and the parameterized model is used to investigate the consequences of hypothetical scenarios of habitat destruction. For each species (the three rows), the six panels give the following information: **a**, a snap-shot of habitat patch occupancy in a fragmented landscape (blue and red dots represent occupied and empty patches, respectively; dot size proportional to habitat patch area); **b**, an area-isolation plot, showing patch occupancy as a function of patch area and connectivity (*S*, Box 2), with the fitted 10, 50 and 90% lines of incidence; **c**, predicted trajectories of patch occupancy during 100 yrs, with no (red line) and two levels of regional stochasticity ($\sigma = 0.4$, black line, and 0.8, blue line; Box 1), as well as the expected fraction of occupied patches (straight line; Box 1); **d**, number of empty patches as a function of the pooled amount of habitat in the remaining patches (horizontal axis) under three forms of habitat destruction: removal of patches within a continuous area (blue), random removal of patches (black), and removal of area (constant fraction) from each patch (red), with filled symbols giving the expected value (Box 1) and open symbols the realized number in a 100-yr simulation (result calculated for the years 51–100); **e**, fraction of metapopulations going extinct in 100 replicate 100-yr simulations in the previous three scenarios of habitat loss, with regional stochasticity set at $\sigma = 0.8$ for the insect and $\sigma = 0.4$ for the other species. Symbols as in **d**, with filled symbols giving the result for landscapes in which the expected metapopulation size was positive; **f**, the time delay in metapopulation response to habitat destruction calculated as described in Box 1. The horizontal axis gives the amount of habitat before the extra habitat destruction, which amounted to 50% for the butterfly and the pika and 20% for the nuthatch (these calculations were made only for random elimination of entire patches and loss of patch area). Open symbols represent cases where the new equilibrium was metapopulation extinction.

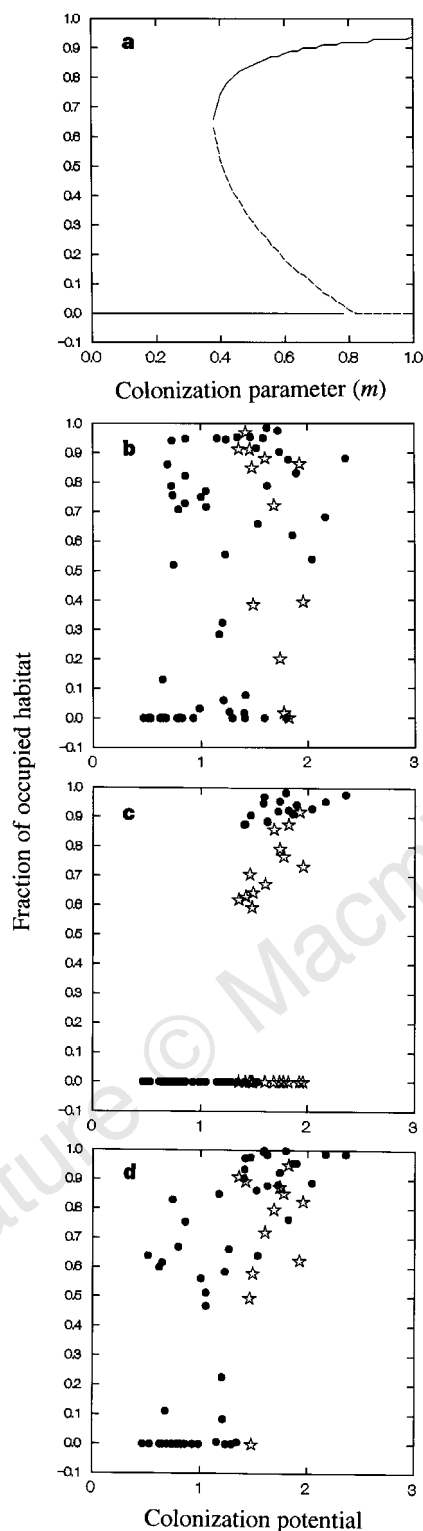


Figure 4 Bifurcation diagrams for the fraction of occupied habitat in metapopulations. **a**, Theoretical result in which the fraction of occupied habitat is plotted against colonization rate parameter in a deterministic structured model⁶⁷. The continuous line represents stable equilibria, the broken line is unstable equilibria. **b**, Empirical result for 66 semi-independent patch networks of the Glanville fritillary butterfly with at least 5 patches⁷⁸ and with the fraction of occupied habitat plotted against the colonization potential described in Box 1. Stars indicate patch networks for which the IFM had two alternative equilibria (Box 1). **c**, Expected fraction of occupied habitat as predicted by the IFM parameterized for the butterfly (Fig. 3b). **d**, One snap-shot from stochastic simulations of the IFM. In this case, a small probability of colonization (0.01) from outside the patch network was assumed for each patch to prevent permanent metapopulation extinction.

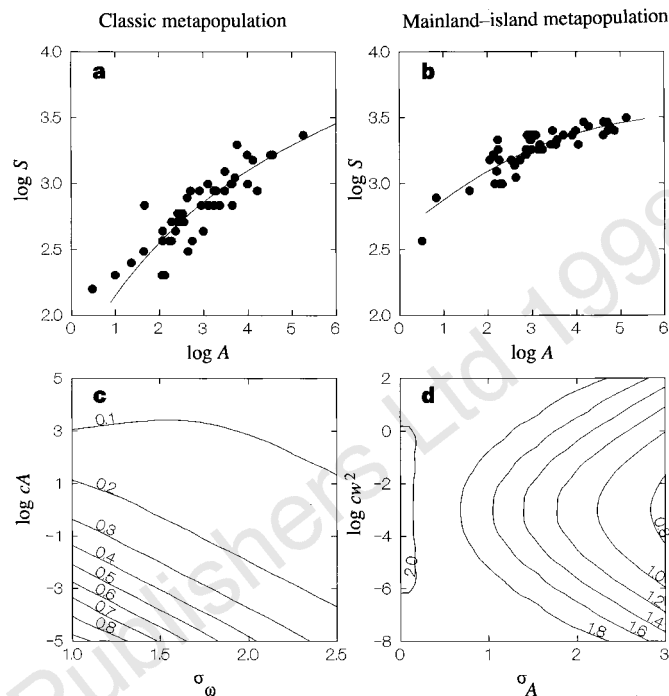


Figure 5 Examples of species-area (SA) curves in classic and mainland-island metapopulations predicted by a metacommunity model of non-interactive species with interspecific differences in population density⁶⁶. **a**, Classic metapopulation; **b**, Mainland-island metapopulation. The lines are the expected relationships, the dots give one stochastic realization, obtained by assigning the presence or absence of species on islands according to the incidences. **c**, **d**, Dependence of the slopes of the SA (**c**) and distribution-abundance (**d**) curves on the isolation of the islands or habitat fragments (increases with decreasing c), average island area (A) and species density (w), and the standard deviations of the respective distributions (σ_A and σ_w) in mainland-island metapopulations. Analogous predictions can be made for classic metapopulations without an external mainland⁶⁶.

which managers operate⁷¹. Further work is needed to extend the single-species metapopulation models to multispecies communities, essentially to merge the spatially explicit and dynamic metapopulation models with non-dynamic site-selection algorithms used in conservation⁷² and with models of habitat connectivity used in landscape ecology⁷³.

If only small fragments of habitat with extinction-prone populations can be preserved, it is desirable to have at least 15–20 fragments located within the migration range of the species, to reduce the probability of metapopulation extinction due to extinction-colonization stochasticity. Managers should absorb the key message of classic metapopulation dynamics: currently unoccupied habitat fragments may be critical for long-term persistence. Optimal spacing of preserved habitat fragments is a compromise between the need to have them located sufficiently close to each other to allow recolonization, but far enough apart to reduce the impact of regional stochasticity. Another means of alleviating regional stochasticity is to include substantial spatial variance in habitat quality among the preserved areas. Habitat quality interacts with stochasticity caused by varying weather conditions, hence a set of dissimilar fragments is unlikely to experience the most unfavourable conditions simultaneously^{74,75}.

One prediction of metapopulation models with great importance for conservation is the time delay with which species are expected to track changes in the structure of fragmented landscapes. We do not know which fraction of currently endangered populations and species are already committed to metapopulation extinction in

their present environments. A real worry is that such 'living dead' populations and species are numerous, especially because the delay in reaching the new equilibrium is particularly long in just those cases that matter most, where the new equilibrium is metapopulation extinction (Fig. 3f).

Finally, it has to be agreed with the critics⁷⁶ who have cautioned against uncritical application of the metapopulation concept and models to conservation. Not all endangered species have the spatial structure of metapopulations, and even if they do, the immediate conservation concern may be elsewhere. Rather than in the conservation of species that are already on the brink of extinction, the metapopulation concept may turn out to be most helpful in the conservation of biodiversity in general in our everyday landscapes. Regionally, many habitats have become so fragmented that isolated populations cannot be expected to last for long, hence long-term persistence can occur only via metapopulation dynamics. □

Ilkka Hanski is in the Department of Ecology and Systematics, Division of Population Biology, PO Box 17 (Arkadiankatu 7), FIN-00014 University of Helsinki, Finland.

1. Nicholson, A. J. The balance of animal populations. *J. Anim. Ecol.* **2**, 132–178 (1933).
2. Andrewartha, H. G. & Birch, L. C. *The Distribution and Abundance of Animals* (University of Chicago Press, Chicago, 1954).
3. Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97–159 (1931).
4. Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. Sixth Int. Congr. Genet.* **1**, 356–366 (1932).
5. Hanski, I. & Gilpin, M. E. *Metapopulation Biology: Ecology, Genetics, and Evolution* (Academic, San Diego, 1997).
6. Tilman, D. & Kareiva, P. *Spatial Ecology* (Princeton University Press, Princeton, 1997).
7. Bascompte, J. & Solé, R. V. *Modeling Spatiotemporal Dynamics in Ecology* (Springer, New York, 1998).
8. Durrett, R. & Levin, S. A. Stochastic spatial models: a user's guide to ecological applications. *Phil. Trans. R. Soc. Lond. B* **343**, 329–350 (1994).
9. Forman, R. T. T. Some general principles of landscape and regional ecology. *Landscape Ecology* **10**, 133–142 (1995).
10. Wiens, J. A. in *Metapopulation Biology* (eds Hanski, I. & Gilpin, M. E.) 3–6. (Academic, San Diego, 1997).
11. Kareiva, P. & Wennergren, U. Connecting landscape patterns to ecosystem and population processes. *Nature* **373**, 299–302 (1995).
12. Wilson, H. B. et al. in *Modeling Spatiotemporal Dynamics in Ecology* (eds Bascompte, J. & Solé, R. V.) 63–82 (Springer, New York, 1998).
13. Nee, S., May, R. M. & Hassell, M. P. in *Metapopulation Biology* (eds Hanski, I. & Gilpin, M. E.) 123–148 (Academic, San Diego, 1997).
14. Harrison, S. in *Large-scale Ecology and Conservation Biology* (eds Edwards, P. J., May, R. M. & Webb, N. R.) 111–128 (Blackwell Scientific, Oxford, 1994).
15. Settle, J. *Species Survival in Fragmented Landscapes* (Kluwer, Dordrecht, 1996).
16. McCullough, D. L. *Metapopulations and Wildlife Conservation* (Island, Washington DC, 1997).
17. Levins, R. Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bull. Entomol. Soc. Am.* **15**, 237–240 (1969).
18. Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* (Oxford University Press, Oxford, 1991).
19. Grenfell, B. & Harwood, J. (Meta)population dynamics of infectious diseases. *Trends Ecol. Evol.* **12**, 395–404 (1997).
20. Bartlett, M. S. Measles periodicity and community size. *J. R. Stat. Soc. A* **120**, 48–70 (1957).
21. Hanski, I. *Metapopulation Ecology* (Oxford University Press, Oxford, in the press).
22. Hanski, I. Single-species metapopulation dynamics: concepts, models and observations. *Biol. J. Linn. Soc.* **42**, 17–38 (1991).
23. Gyllenberg, M., Hanski, I. & Hastings, A. in *Metapopulation Biology* (eds Hanski, I. & Gilpin, M. E.) 93–122 (Academic, San Diego, 1997).
24. Hanski, I. Patch-occupancy dynamics in fragmented landscapes. *Trends Ecol. Evol.* **9**, 131–135 (1994).
25. Hanski, I. in *Metapopulation Biology* (eds Hanski, I. & Gilpin, M. E.) 69–92 (Academic, San Diego, 1997).
26. Hanski, I., Moilanen, A. & Gyllenberg, M. Minimum viable metapopulation size. *Am. Nat.* **147**, 527–541 (1996).
27. Chesson, P. in *Modeling Spatiotemporal Dynamics in Ecology* (eds Bascompte, J. & Solé, R. V.) 151–166 (Springer, New York, 1998).
28. Thomas, C. D. & Hanski, I. in *Metapopulation Biology* (eds Hanski, I. & Gilpin, M. E.) 359–386 (Academic, San Diego, 1997).
29. Saccheri, I. J. et al. Inbreeding and extinction in a butterfly metapopulation. *Nature* **392**, 491–494 (1998).
30. Lande, R. Genetics and demography in biological conservation. *Science* **241**, 1455–1460 (1988).
31. Lande, R. Risks of population extinction from demographic and environmental stochasticity and random catastrophes. *Am. Nat.* **142**, 911–927 (1993).
32. Foley, P. Predicting extinction times from environmental stochasticity and carrying capacity. *Cons. Biol.* **8**, 124–137 (1994).
33. Royama, T. *Analytical Population Dynamics* (Chapman & Hall, London, 1992).
34. Turchin, P. in *Population Dynamics: New Approaches and Synthesis* (eds Cappuccino, N. & Price, P.) 19–40 (Academic, London, 1995).
35. den Boer, P. J. Seeing the tree for the wood: random walks or bounded fluctuations of population size? *Oecologia* **86**, 484–491 (1991).
36. Tilman, D. et al. Habitat destruction and the extinction debt. *Nature* **371**, 65–66 (1994).
37. Hastings, A. & Harrison, S. Metapopulation dynamics and genetics. *Annu. Rev. Ecol. Syst.* **25**, 167–188 (1994).
38. Hassell, M. P., Comins, H. N. & May, R. M. Species coexistence and self-organizing spatial dynamics. *Nature* **370**, 290–292 (1994).
39. Barton, N. H. & Whitlock, M. C. in *Metapopulation Biology* (eds Hanski, I. & Gilpin, M. E.) 183–214 (Academic, San Diego, 1997).
40. Hedrick, P. W. & Gilpin, M. E. in *Metapopulation Biology* (eds Hanski, I. & Gilpin, M. E.) 166–182 (Academic, San Diego, 1997).
41. Olivieri, I. & Gouyon, P.-H. in *Metapopulation Biology* (eds Hanski, I. & Gilpin, M. E.) 3–324 (Academic, San Diego, 1997).
42. Murphy, D. D., Freas, K. S. & Weiss, S. B. An environment-metapopulation approach to the conservation of an endangered invertebrate. *Cons. Biol.* **4**, 41–51 (1990).
43. Coyne, J. A., Barton, N. H. & Turelli, M. A critique of Wright's shifting balance theory of evolution. *Evolution* **51**, 643–671 (1997).
44. Leimar, O. & Nordberg, U. Metapopulation extinction and genetic variation in dispersal-related traits. *Oikos* **80**, 448–458 (1997).
45. Wahlberg, N., Moilanen, A. & Hanski, I. Predicting the occurrence of endangered species in fragmented landscapes. *Science* **273**, 1536–1538 (1996).
46. Ims, R. A. & Yoccoz, N. G. in *Metapopulation Biology* (eds Hanski, I. & Gilpin, M. E.) 247–266 (Academic, San Diego, 1997).
47. Turchin, P. *Quantitative Analysis of Movement: Measuring and Modelling Population Redistribution in Animals and Plants* (Sinauer, Sunderland, MA, 1998).
48. Hanski, I. A practical model of metapopulation dynamics. *J. Anim. Ecol.* **63**, 151–162 (1994).
49. ter Braak, C. J. F., Hanski, I. A. & Verboom, J. in *Modeling Spatiotemporal Dynamics in Ecology* (eds Bascompte, J. & Solé, R. V.) 167–188 (Springer, New York, 1998).
50. Moilanen, A. Patch occupancy models of metapopulation dynamics: efficient parameter estimation with implicit statistical inference. *Ecology* (in the press).
51. Moilanen, A., Smith, A. T. & Hanski, I. A. Long-term dynamics in a metapopulation of the American pika. *Am. Nat.* **152**, 530–542 (1998).
52. Hanski, I. Inferences from ecological incidence functions. *Am. Nat.* **139**, 657–662 (1992).
53. Pacala, S. W. & Levin, S. A. in *Spatial Ecology* (eds Tilman, D. & Kareiva, P.) 204–232 (Princeton University Press, Princeton, 1997).
54. Cain, M. L. et al. Neighbourhood models of clonal growth in the white clover *Trifolium repens*. *Am. Nat.* **145**, 888–917 (1995).
55. Turchin, P. et al. in *Modeling Spatiotemporal Dynamics in Ecology* (eds Bascompte, J. & Solé, R. V.) 199–213 (Springer, New York, 1998).
56. Maron, J. L. & Harrison, S. Spatial pattern formation in an insect host-parasitoid system. *Science* **278**, 1619–1621 (1998).
57. Hanski, I. & Gyllenberg, M. Two general metapopulation models and the core-satellite species hypothesis. *Am. Nat.* **142**, 17–41 (1993).
58. Temple, S. A. The problem of avian extinctions. *Curr. Ornithol.* **3**, 453–485 (1986).
59. Akçakaya, H. R. & Ferson, S. *RAMAS/Space User Manual: Spatially Structured Population Models for Conservation Biology* (Applied Biomathematics, New York, 1992).
60. Possingham, H. P. & Noble, I. R. in *Research Consultancy for the Research Assessment Commission, Forest and Timber Inquiry* (Canberra, 1991).
61. Sjögren-Gulve, P. & Ray, C. in *Metapopulations and Wildlife Conservation* (ed McCullough, D. R.) 111–138 (Island, Washington DC, 1997).
62. Andrén, H. in *Mosaic Landscapes and Ecological Processes* (eds Hansson, L., Fahrig, L. & Merriam, G.) 225–255 (Chapman & Hall, London, 1995).
63. Nee, S. & May, R. M. Dynamics of metapopulations: Habitat destruction and competitive coexistence. *J. Anim. Ecol.* **61**, 37–40 (1992).
64. Caswell, H. & Cohen, J. E. Disturbance, interspecific interaction and diversity in metapopulations. *Biol. J. Linn. Soc.* **42**, 193–218 (1991).
65. Holt, R. D. in *Metapopulation Biology* (eds Hanski, I. & Gilpin, M. E.) 149–165 (Academic, San Diego, 1997).
66. Hanski, I. & Gyllenberg, M. Uniting two general patterns in the distribution of species. *Science* **275**, 397–400 (1997).
67. MacArthur, R. H. & Wilson, E. O. *The Theory of Island Biogeography* (Princeton University Press, Princeton, 1967).
68. Lawton, J. H. Range, population abundance and conservation. *Trends Ecol. Evol.* **8**, 409–413 (1993).
69. Moilanen, A. & Hanski, I. Metapopulation dynamics: effects of habitat quality and landscape structure. *Ecology* **79**, 2503–2515 (1998).
70. Hanski, I. & Simberloff, D. in *Metapopulation Biology* (eds Hanski, I. & Gilpin, M. E.) 5–26 (Academic, San Diego, 1997).
71. Thomas, C. D. in *Aspects of the Genesis and Maintenance of Biological Diversity* (eds Hochberg, M. E., Clobert, J. & Barbault, R.) 292–307 (Oxford University Press, Oxford, 1996).
72. Pressey, R. L., Possingham, H. P. & Day, J. R. Effectiveness of alternative heuristic algorithms for identifying indicative minimum requirements for conservation reserves. *Biol. Cons.* **80**, 207–219 (1997).
73. With, K. A., Gardner, R. H. & Turner, M. G. Landscape connectivity and population distributions in heterogeneous landscapes. *Oikos* **78**, 151–169 (1997).
74. Weiss, S. J., Murphy, D. D. & White, R. R. Sun, slope and butterflies: topographic determinants of habitat quality in *Euphydryas editha*. *Ecology* **69**, 1486–1496 (1988).
75. Kindvall, O. The impact of extreme weather on habitat preference and survival in a metapopulation of the bush cricket *Metrioptera bicolor* in Sweden. *Biol. Cons.* **73**, 51–58 (1995).
76. Doak, D. F. & Mills, L. S. A useful role for theory in conservation. *Ecology* **75**, 615–626 (1994).
77. Gyllenberg, M. & Hanski, I. Habitat deterioration, habitat destruction and metapopulation persistence in a heterogeneous landscape. *Theor. Pop. Biol.* **52**, 198–215 (1997).
78. Hanski, I. et al. Multiple equilibria in metapopulation dynamics. *Nature* **377**, 618–621 (1995).

Acknowledgements. I thank S. Harrison, M. Hassell, T. Ives, J. Lawton, A. Moilanen, S. van Nouhuys, B. O'Hara, A. Ruina, I. Saccheri, M. Singer, C. Thomas, P. Turchin and N. Wahlberg for comments on the manuscript, and J. Verboom and A. Smith for supplying the nuthatch and pika data for Fig. 3. My metapopulation research has been supported by the Academy of Finland.

Evidence against stellar chromospheric origin of Galactic cosmic rays

Andrew J. Westphal*, P. Buford Price*, Benjamin A. Weaver* & Vladimir G. Afanasiev*†

* Department of Physics and Space Sciences Laboratory, University of California at Berkeley, Berkeley, California 94720, USA

Interstellar space is filled with a gas of relativistic ions and electrons—the Galactic cosmic rays. These energetic particles tie interstellar gas to ambient magnetic fields by ionizing the component molecules and atoms, and so play a role in stabilizing molecular clouds against collapse¹ and regulating the collapse of protostellar clouds². The observed energy spectrum of cosmic rays up to $\geq 10^{15}$ eV is consistent with their acceleration by supernova shock waves³, but the original source of cosmic-ray nuclei remains unclear. There has been a widely held belief that the source consists of a solar-like ionized medium⁴, probably the chromospheres of late-type Sun-like stars⁵. This model predicts an overabundance of easily ionized elements. Here we show that lead, which is easily ionized, is underabundant in the Galactic cosmic rays in contradiction with this model. Rather, our measurements are consistent with two other possible models: one in which the nuclei originate in interstellar gas, and in entire grains accelerated to about one per cent of the speed of light by supernova shock waves^{6,7}; and another in which the cosmic rays contain an admixture of an exotic, freshly synthesized component⁸, probably originating in neutrino-driven winds from newly born neutron stars⁹.

We used the Trek large collector—an array of 150 stacks of BP-1 (ref. 10) barium phosphate glass track-etch detectors¹¹—deployed on the Russian space station Mir from 1991 to 1995. The stacks were deployed in a 1.2 m² array on the outside surface of Mir. Each stack consisted of 16 glass detectors, each measuring 9 cm × 9 cm × 1.5 mm. One-third of the stacks were returned to Earth after 2.5 years in a dedicated re-entry capsule. The remaining two-thirds were recovered by the space shuttle Atlantis during the second shuttle–Mir rendezvous after a 4.2-year exposure. After recovery, each stack was exposed uniformly to a low-fluence 10.6 GeV per nucleon Au calibration beam at the BNL Alternating Gradient Synchrotron. The second batch of stacks returned to Earth was exposed to beams at two different exposure angles, which enabled cross-checking of measurements and verification of zenith-angle independence of charge measurement. The detectors were then annealed to correct for differential track-fading on orbit. Three pairs of adjacent detectors were selected from the top, middle and bottom of the stacks for analysis of Pt- and Pb-group elements. The remaining ten detectors from each stack were archived in cold storage for analysis of actinides and elements with atomic number $Z < 70$, which require different analysis techniques.

Heavily ionizing particles which penetrate the glass track-etch detector leave a track which is chemically more reactive than the bulk of the glass¹¹. When the detectors are placed in an appropriate etchant, the tracks are etched at a rate v_T which is larger than the general etch rate v_G . The etching process thus creates a conical etch-pit, with an elliptical mouth, at each surface transversed by the particle. The detector signal is the reduced track-etch rate $s \equiv v_T/v_G$. BP-1 exhibits remarkably high resolution (see, for example, refs 12, 13), and has the advantage that the dynamic range may be ‘tuned’ by choice of etchant¹³. For the analysis of the Pt–Pb region of the

periodic table, we etched the detectors for 72 hours in 49% hydrofluoric acid at room temperature, removing $\sim 50 \mu\text{m}$ from each surface. This thickness removed from each detector was measured to better than 1%. The cosmic-ray tracks were located by both automated and manual microscopic scanning. We found 255 ions which penetrated at least half of the stack. The scanning efficiency was measured to be $>90\%$. Measurements were made of the etch-pit sizes and particle trajectories, and Au calibration tracks in a 2 cm² area around each cosmic-ray etch-pit were measured automatically. This fully local calibration is a unique feature of the Trek experiment. Additional details of the processing and measurement of the glass detectors will be given elsewhere¹⁴.

The detector signal s versus flight distance for each cosmic-ray track was then computed from these measurements. These data were used to derive a nuclear charge Z and kinetic energy for each particle, using the empirical response of the detector and a model of relativistic heavy-ion slowing. We verified the experimental technique by a variety of methods¹⁴. The detector response was locally corrected using calibration measurements. The charge resolution σ_z using this technique due to known sources of dispersion is (0.39–0.45) e , with at most 12 measurements and this set of particle energies and trajectories. This charge resolution is determined mainly by intrinsic fluctuations in signal, and by limited stack thickness, as energy must simultaneously be determined with

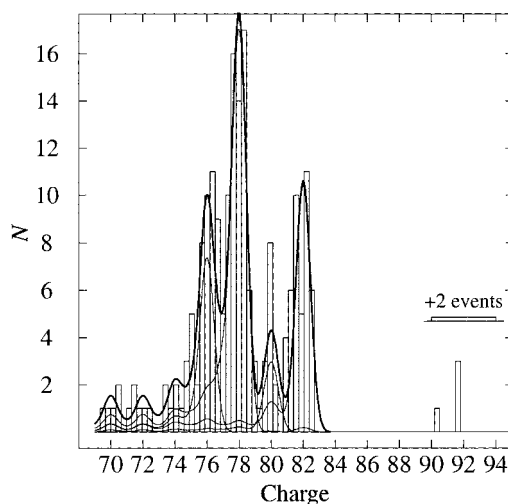


Figure 1 Histogram of measured charges for GCR ions with kinetic energies >0.9 GeV per nucleon. The charge resolution is $\sim 0.45e$, roughly a three-fold improvement in resolution over pioneering measurements by HEAO²³, Ariel¹⁹ and UHCRES²⁰. The charge peaks were shifted to lower values by $0.5e$ relative to the calibration beam due to thermal fading of the tracks on orbit. We also found six actinides ($Z \geq 89$), all of which had an apparent charge $Z > 90$. In the original analysis the signal for these events was nearly saturated, so the charge resolution was only sufficient to identify the particles as actinides. We re-analysed these six tracks, using six unetched sheets from each of the stacks, in a NaOH etch which gives an unsaturated signal at the actinides¹³, but which does not allow the use of the Au calibration. Two of the six events were not analysable in NaOH due to large zenith angle and nuclear fragmentation, respectively. To be conservative, we treat these events as ambiguous for the purpose of calculating systematic error. We show individual single-charge spread functions (see text) of the best-fit source composition in the Pt–Pb region, and their sum, uncorrected for detector acceptance, assuming 3.0 GeV per nucleon propagation energy, 0 g cm^{−2} truncation length, and $\sigma_z = 0.45e$. The source composition was determined by maximum likelihood ($\chi^2 = 39.2$ with 39 d.o.f.). Most Ir decays to Pt and most Tl to Pb immediately after acceleration due to bound-state β -decay²⁶, so the observed abundances of each pair are highly correlated; this causes a numerical instability in the derivation of source abundances. As their respective atomic properties are nearly identical, we combined the pairs in solar proportions in deriving source abundances, then subtracted Ir and Tl to derive source abundances of Pt and Pb.

† Present address: Genesys Corporation, San Francisco, California 94103, USA.

charge by measurement of slowing across the stack. BP-1 achieves a dispersion in measurement of instantaneous charge change of $\sim 0.15e$ per etch-pit, so integral changes in apparent charge due to occasional nuclear fragmentation or temporary capture of an electron into an atomic orbital were detected and taken into account. In Fig. 1, we show a histogram of measured charge, for 192 particles with kinetic energy > 0.9 GeV per nucleon at the top of the detector. The median kinetic energy of these particles was 2.4 GeV per nucleon.

The measurements of elemental abundances at the detector, corrected for detector acceptance but not for propagation effects all with respect to the Pt-group elements ($75 \leq Z \leq 79$) are: ${}_{76}\text{Os} = 0.38^{+0.06}_{-0.05}$ (statistical) $^{+0.07}_{-0.03}$ (systematic), ${}_{77}\text{Ir} < 0.21$, ${}_{78}\text{Pt} =$

$0.48^{+0.04}_{-0.06}$ (stat.) $^{+0.02}_{-0.01}$ (syst.), ${}_{79}\text{Au} < 0.05$, ${}_{80}\text{Hg} = 0.12^{+0.03}_{-0.04}$ (stat.) $^{+0.01}_{-0.02}$ (syst.), ${}_{82}\text{Pb} = 0.29^{+0.05}_{-0.06}$ (stat.) $^{+0.01}_{-0.02}$ (syst.) and actinides = $0.034^{+0.016}_{-0.015}$ (stat.) $^{+0.001}_{-0.011}$ (syst.). Here we have assumed that the abundances of the rare elements Re, Tl and Bi are negligible, but have included their presence in solar proportions as systematic errors in Os, Hg and Pb. No particles measured with good charge resolution were found in the actinide gap ($83 < Z \leq 89$) or with $Z > 92$. This null result may be used to set upper limits on the abundances in Galactic cosmic rays (GCRs) of hypothetical long-lived superheavy elements¹⁵ and exotic, heavily ionizing particles.

During flight, GCR nuclei suffer nuclear fragmentation due to collisions with nuclei in the interstellar medium. The cross-sections for nuclear fragmentation by ultraheavy nuclei have recently been measured in the energy interval 1–10 GeV per nucleon by the UHIC collaboration¹⁶. Using these cross-sections, we calculated the abundances of nuclear fragments for hypothetical pure beams of $Z = 70$ –92, due to propagation in a leaky box¹⁷. Our calculations agree with those of C. J. Waddington (personal communication). For each element, we calculated a single-charge spread function, that is, a distribution of observed charge due to that element present at the GCR source, taking into account interstellar propagation and detector resolution. We then derived abundances at the GCR source by fitting our unbinned charge measurements to elemental single-charge spread functions, corrected for detector acceptance, using maximum likelihood.

It has been recognized for some time that the GCR elemental and isotopic abundances are generally similar to Solar System values, and that easily ionizable elements are systematically overabundant⁴. A detailed physical model was proposed by Meyer⁵, in which stellar flares accelerate ions from the ionized gas in chromospheres of Sun-like stars to modest energies before subsequent acceleration to high energies by supernova shocks. That just such a fractionation based on ionizability is observed in solar energetic particle acceleration¹⁸ from the solar chromosphere gives additional support to this model. Our measurement of the Pb abundance is a factor of ~ 3 smaller than the abundance predicted by a Sun-like or protosolar source with acceleration related to first ionization potential (FIP) or no preferential acceleration ("lg-fip" and "lg-no" in Fig. 2), thus strongly disfavouring any model in which GCR nuclei originate in an ionized gas of local Galactic composition. The HEAO-3⁸, Ariel¹⁹ and UHCRES²⁰ instruments also reported low Pb abundance, but with insufficient charge resolution to resolve even- Z elements in the Pt–Pb region.

Our data are consistent with a model ("lg-vol" in Fig. 2) originally proposed by Epstein²¹ and Cesarsky and Bibring²², and recently refined and extended by Meyer, Drury and Ellison^{6,7}, in which the GCR source consists of interstellar gas and ions sputtered from grains accelerated to ~ 0.1 MeV per nucleon by supernova shocks. This model predicts that volatile elements, rather than those with large FIP, are depleted in the GCR, as volatile elements do not condense into grains. Most volatile elements have high FIP, and most refractory elements have low FIP. One exception to this rule is Pb, which has low FIP but is volatile, so a measurement of Pb can distinguish between FIP and volatility as the correct organizing atomic property. The observed depletion of Pb is consistent with a source of local Galactic composition with volatility-biased preferential acceleration, in agreement with the gas and grain acceleration model of Meyer, Drury and Ellison^{6,7}.

Most elements with $Z > 32$ are believed to be built up from lighter seed nuclei through neutron addition and β -decay. Nuclei with closed neutron shells ($N = 50, 82, 126$) serve as bottlenecks for this process, and are overproduced. If the neutron absorption rate is slower than typical β -decay lifetimes, abundance peaks occur at near ${}_{40}\text{Zr}$, ${}_{56}\text{Ba}$ and ${}_{82}\text{Pb}$; if faster, peaks occur near ${}_{36}\text{Kr}$, ${}_{50}\text{Sn}$ and ${}_{78}\text{Pt}$. The site of the slow ('s-') process is probably asymptotic giant branch stars; and site of the rapid ('r-') process is probably a neutrino-driven wind blown from nascent neutron stars⁹. Our

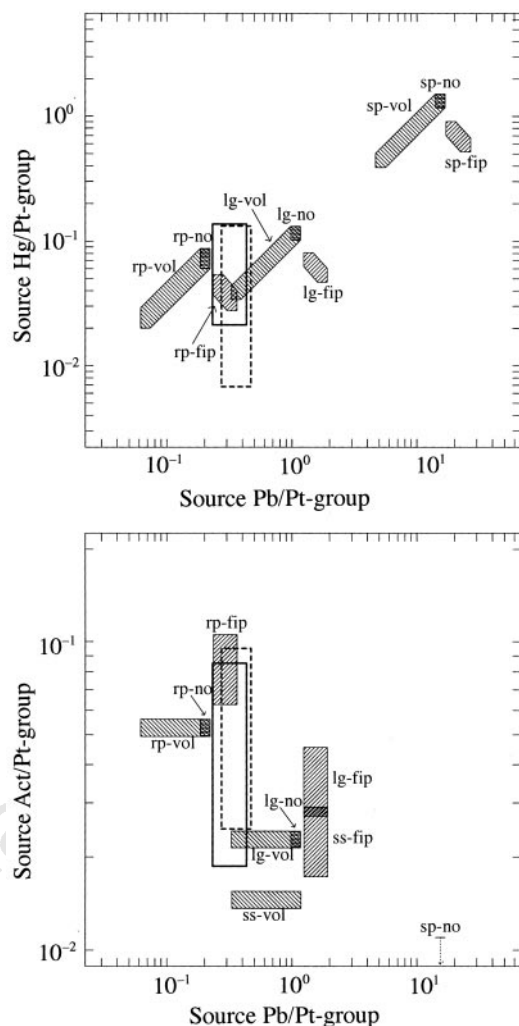


Figure 2 Derived GCR source abundances of Hg, Pb and actinides with respect to the Pt-group elements ($75 \leq Z \leq 79$). Two assumptions have been made for GCR propagation: no truncation (solid box), and 1 g cm^{-2} truncation (dashed box). We compare our measurements to hypothetical sources of current solar ("ss"²⁷ composition; local Galactic ("lg") composition, expected to be nearly identical to protosolar values^{27,28}, pure fresh r-process ("rp") composition^{29,30}, and pure s-process ("sp") composition³⁰; and with no fractionation at the source (suffix "no"), FIP-biased fractionation (suffix "fip")⁵, and volatility-biased fractionation ("vol")^{6,7}. The confidence limits of these measurements were determined by analysing 100 random distributions based on the measured composition. We repeated this analysis for a range of propagation energies (2.0–3.0 GeV per nucleon) and several choices of charge resolution within a conservative range, (0.39–0.50)e. The dispersion of the Pb peak, 0.49e, may indicate a slight worsening of charge resolution at $Z = 82$, although the χ^2 of the overall fit is acceptable for $\sigma_c = 0.45e$. Our measurement limits consist of the extrema of the 1σ limits from each of these analyses.

measurements strongly rule out a source dominated by material synthesized by the s-process ("sp-" models in Fig. 2), on the basis of Hg and Pb abundances and the presence of actinides, which are not produced by the s-process at all (Fig. 2). Using HEAO data, Binns *et al.*⁸ did not rule out an s-process source based on Sn, Te and Ba abundances. In an analysis of measurements of GCR composition using the HEAO-3²³ and Ariel-6¹⁹ instruments, Binns *et al.*⁸ have proposed an empirical model in which GCRs with $Z < 60$ have a solar-like composition, but those with $Z > 60$ are dramatically enhanced in r-process material. Our data are consistent with a source consisting of freshly synthesized r-process material with FIP-ordered preferential acceleration ("rp-fip" in Fig. 2). Although an r-process enhancement might be expected in supernova ejecta, no detailed model has been proposed which gives such an enhancement only for $Z > 60$. However, there is some indication from presolar abundances of ^{129}I and ^{182}Hf of an r-process component which preferentially synthesizes nuclei with atomic weight $A > 140$ ($Z \geq 58$)²⁴.

Ramaty, Lingenfelter and Kozlovsky have recently suggested that GCRs originate in grains condensed from fresh supernova ejecta²⁵. Such a source would be enriched in r-process material and refractory elements (qualitatively similar to "rp-vol" in Fig. 2). A pure r-process source with ≥ 1.5 acceleration bias is inconsistent with the observed Pb abundance.

The interstellar grain plus gas model of Meyer, Drury and Ellison^{6,7}, and the empirical r-process enhancement model of Binns *et al.*⁸ discussed above predict very similar Hg and Pb abundances (Fig. 2), but they could be distinguished by a more accurate measurement of the overall actinide abundances. Also, freshly synthesized r-process material, predicted by the Binns *et al.* model and by Ramaty, Lingenfelter and Kozlovsky²⁵, would necessarily be young, and the unstable actinides (Th, U, Pu, Cm) could serve as 'clocks' to measure the age of the nuclei. The actinides are quite rare, so an exposure of $\sim 100 \text{ m}^2 \text{ yr sr}$ is required to accurately measure age. ECCO, a detector similar to Trek currently under study for deployment on the International Space Station, is expected to have sufficient collecting power and charge resolution to achieve this goal. □

Received 15 June; accepted 28 July 1998.

- Zweibel, E. G. & Heiles, C. Magnetic fields in galaxies and beyond. *Nature* **385**, 131–136 (1997).
- McKee, C. F., Zweibel, E. G., Heiles, C. & Goodman, A. A. in *Protostars and Planets III* (eds Levy, E. H. & Lunine, J. I.) 327–366 (Univ. Arizona Press, Tucson, 1993).
- Gaissner, T. K. *Cosmic Rays and Particle Physics* (Cambridge Univ. Press, 1990).
- Webber, W. R. New experimental data and what it tells us about the sources and acceleration of cosmic rays. *Space Sci. Rev.* **81**, 107–142 (1997).
- Meyer, J.-P. Solar-stellar outer atmospheres and energetic particles, and galactic cosmic rays. *Astrophys. J. Suppl.* **57**, 173–204 (1985).
- Meyer, J.-P., Drury, L. O'C. & Ellison, D. C. Galactic cosmic rays from supernova remnants. I. A cosmic-ray composition controlled by volatility and mass-to-charge ratio. *Astrophys. J.* **487**, 182–196 (1997).
- Ellison, D. C., Meyer, J.-P. & Drury, L. O'C. Galactic cosmic rays from supernova remnants. II. Shock acceleration of gas and dust. *Astrophys. J.* **487**, 197–217 (1997).
- Binns, W. R. *et al.* in *Cosmic Abundances of Matter* (ed. Waddington, C. J.) 147–167 (Am. Inst. Physics, New York, 1989).
- Meyer, B. S. The r-, s-, and p-processes. *Annu. Rev. Astron. Astrophys.* **32**, 153–190 (1994).
- Wang, S.-C. *et al.* Phosphate glass detectors with high sensitivity to nuclear particles. *Nucl. Instrum. Meth. B* **35**, 43–49 (1988).
- Fleischer, R. L., Price, P. B. & Walker, R. M. *Nuclear Tracks in Solids* (Univ. California Press, Berkeley, 1975).
- Westphal, A. J. & He, Y. D. Measurement of cross sections for electron capture and stripping by highly relativistic ions. *Phys. Rev. Lett.* **71**, 1160–1163 (1993).
- Westphal, A. J., Price, P. B. & Snowden-Ifft, D. P. Upper limit on the cross section for nuclear charge pickup by relativistic uranium ions. *Phys. Rev. C* **45**, 2423–2426 (1992).
- Weaver, B. A. *et al.* Performance of the ultraheavy collector of the Trek experiment. *Nucl. Instrum. Meth.* (submitted).
- Möller, P. & Nix, J. R. Stability and decay of nuclei at the end of the periodic system. *J. Phys. G* **20**, 1681–1747 (1994).
- Geer, L. Y. *et al.* Charge-changing fragmentation of 10.6 GeV/nucleon ^{197}Au nuclei. *Phys. Rev. C* **52**, 334–345 (1995).
- Clinton, R. R. & Waddington, C. J. Dependence of the interstellar propagation of ultraheavy cosmic-ray nuclei on various parameters. *Astrophys. J.* **403**, 644–657 (1993).
- Reames, D. V. Coronal abundances determined from energetic particles. *Adv. Space Res.* **15**, 41–51 (1995).
- Fowler, P. H. *et al.* Ariel 6 measurements of the fluxes of ultraheavy cosmic rays. *Astrophys. J.* **314**, 739–746 (1987).
- O'Sullivan, D., Thompson, A., Keane, A. J., Drury, L. O'C. & Wenzel, K.-P. Investigation of $Z \geq 70$ cosmic ray nuclei on the LDEF mission. *Radiat. Meas.* **26**, 889–892 (1996).

- Epstein, R. I. The acceleration of interstellar grains and the composition of the cosmic rays. *Mon. Not. R. Astron. Soc.* **193**, 723–729 (1980).
- Cesarsky, C. J. & Bibering, J. P. in *Origin of Cosmic Rays* (eds Setti, G., Spada, G. & Wolfendale, A. W.) 361–362 (IAU Symp. 94, Reidel, Dordrecht, 1980).
- Binns, W. R. *et al.* Abundances of ultraheavy elements in the cosmic radiation: results from HEAO 3. *Astrophys. J.* **346**, 997–1009 (1989).
- Wasserburg, G. J., Busso, M. & Gallino, R. Abundances of actinides and short-lived nonactinides in the interstellar medium: diverse supernova sources for the r-processes. *Astrophys. J.* **466**, L109–L113 (1996).
- Ramaty, R., Kozlovsky, B. & Lingenfelter, R. Cosmic rays, nuclear gamma rays and the origin of the light elements. *Phys. Today* **51**, 30–35 (1998).
- Takahashi, K., Boyd, R. N., Mathews, G. J. & Yokoi, K. Bound-state beta decay of highly ionized atoms. *Phys. Rev. C* **36**, 1522–1528 (1987).
- Anders, E. & Grevesse, N. Abundances of the elements: meteoritic and solar. *Geochim. Cosmochim. Acta* **53**, 197–214 (1989).
- Pfeiffer, K., Kratz, K.-L. & Thielemann, F.-K. Analysis of the solar-system r-process abundance pattern with the new ETFSI-Q mass formula. *Z. Phys. A* **357**, 235–238 (1997).
- Käppeler, F., Beer, H. & Wisshak, K. s-process nucleosynthesis—nuclear physics and the classical model. *Rep. Prog. Phys.* **52**, 945–1013 (1989).
- Beer, H., Corvi, F. & Mutti, P. Neutron capture of the bottleneck isotopes ^{138}Ba and ^{208}Pb , s-process studies, and the r-process abundance distribution. *Astrophys. J.* **474**, 843–861 (1997).

Acknowledgements. We thank D. O'Sullivan, M. Solarz, V. Akimov, Y. He, the staff at the AGS, and the crews of Mir and STS-74 for their assistance. We also thank the UHIC collaboration for providing measured cross-sections, and C. J. Waddington, B. S. Mayer and J.-P. Meyer for calculations and discussions.

Correspondence and requests for materials should be addressed to A.W. (e-mail: westphal@dilbert.berkeley.edu).

Complete quantum teleportation using nuclear magnetic resonance

M. A. Nielsen^{*†}, E. Knill[‡] & R. Laflamme^{*}

^{*} *Theoretical Astrophysics T-6, MS B-288, and † Computer Research and Applications CIC-3, MS B-265, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*

[‡] *Department of Physics and Astronomy, University of New Mexico, Albuquerque, New Mexico 87131-1131, USA*

Quantum-mechanical systems have information processing capabilities^{1,2} that are not possible with classical devices. One example is quantum teleportation³, in which the quantum state of a system is transported from one location to another without moving through the intervening space. But although partial implementations^{4,5} of quantum teleportation over macroscopic distances have been achieved using optical systems, the final stage of the teleportation procedure—which allows the complete recovery of the original state—was omitted. Here we report an experimental implementation of full quantum teleportation over interatomic distances using liquid-state nuclear magnetic resonance. We achieve teleportation of the quantum state of a carbon nucleus to a hydrogen nucleus in molecules of trichloroethylene, by exploiting natural phase decoherence of the carbon nuclei. Such a teleportation scheme may be used as a subroutine in larger quantum computations, or for quantum communication.

In classical physics, an object can be teleported, in principle, by performing a measurement to characterize completely the properties of the object. That information can then be sent to another location, and the object reconstructed. Does this provide a complete reconstruction of the original object? No: all physical systems are ultimately quantum mechanical, and quantum mechanics tells us that it is impossible to determine completely the state of an unknown quantum system, making it impossible to use the classical measurement procedure to move a quantum system from one location to another.

Bennett *et al.*³ have suggested a procedure for teleporting quantum states. Quantum teleportation may be described abstractly in terms of two parties, Alice and Bob. Alice has in her possession an unknown state $|\Psi\rangle = \alpha|0\rangle + \beta|1\rangle$ of a single quantum bit (qubit)—a two-level quantum system. The goal of teleportation is to trans-

port the state of that qubit to Bob. In addition, Alice and Bob each possess one qubit of a two-qubit entangled state;

$$|\Psi\rangle_A(|0\rangle_A|0\rangle_B + |1\rangle_A|1\rangle_B) \quad (1)$$

where subscripts A are used to denote Alice's systems, and subscripts B to denote Bob's systems. Here and throughout we omit overall normalization factors from our equations.

This state can be rewritten in the Bell basis ($(|00\rangle \pm |11\rangle)$, $(|01\rangle \pm |10\rangle)$) for the first two qubits and a conditional unitary transformation of the state $|\Psi\rangle$ for the last one, that is;

$$\begin{aligned} &(|00\rangle + |11\rangle)|\Psi\rangle + (|00\rangle - |11\rangle)\sigma_z|\Psi\rangle \\ &+ (|01\rangle + |10\rangle)\sigma_x|\Psi\rangle + (|01\rangle - |10\rangle)(-i\sigma_y|\Psi\rangle) \end{aligned} \quad (2)$$

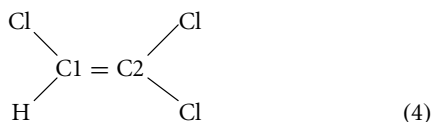
where σ_x , σ_y , σ_z are the Pauli sigma operators⁶, in the $|0\rangle$, $|1\rangle$ basis. A measurement is performed on Alice's qubits in the Bell basis. Conditional on these measurement outcomes it may be verified from equation (2) that Bob's respective states are:

$$|\Psi\rangle; \quad \sigma_z|\Psi\rangle; \quad \sigma_x|\Psi\rangle; \quad -i\sigma_y|\Psi\rangle \quad (3)$$

Alice sends the outcome of her measurement to Bob, who can then recover the original state $|\Psi\rangle$ by applying the appropriate unitary transformation I , σ_z , σ_x or $i\sigma_y$, conditional on Alice's measurement outcome. We note that the quantum state transmission has not been accomplished faster than light because Bob must wait for Alice's measurement result to arrive before he can recover the quantum state.

Recent demonstrations of quantum teleportation^{4,5} omitted the final stage of teleportation, the unitary operators applied by Bob conditional on the result of Alice's measurement. This prevents complete recovery of the original state. Instead, the earlier experiments relied on classical post-processing of the data after completion of the experiment to check that the results were consistent with what one would expect if the conditional operations had, in fact, been performed. Our experiment implements the full teleportation operation. The most important implication of the inclusion of this extra stage is that our teleportation procedure can, in principle, be used as a subroutine in the performance of other quantum information processing tasks. Teleportation as a subroutine is important in potential applications to quantum computation and communication^{7,8}, although in our experimental system, moving Alice's qubit to Bob may be accomplished more efficiently by techniques other than teleportation.

Our implementation of teleportation is performed using liquid-state nuclear magnetic resonance (NMR), applied to an ensemble of molecules of labelled trichloroethylene (TCE). The structure of the TCE molecule may be depicted as:



To perform teleportation we make use of the hydrogen nucleus (H), and the two ^{13}C nuclei (C1 and C2), teleporting the state of C2 to H. Figure 1a illustrates the teleportation process we used. The circuit has three inputs, which we will refer to as the data (C2), ancilla (C1) and target (H) qubits. The goal of the circuit is to teleport the state of the data qubit so that it ends up on the target qubit. We are therefore only teleporting the qubit a few ångströms, making this a demonstration of the method of teleportation, rather than a practical means for transmitting qubits over long distances.

State preparation is done in our experiment using the gradient-pulse techniques described by Cory *et al.*⁹, and phase cycling^{10,11}. The unitary operations performed during teleportation may be implemented in a straightforward manner in NMR, using non-selective radio frequency (r.f.) pulses tuned to the Larmor frequencies of the nuclear spins, and delays allowing entanglement to form through

the interaction of neighbouring nuclei^{9,12}. Other demonstrations of quantum information processing with three qubits using NMR are described in refs 13–15, and with two qubits in refs 16–19.

An innovation in our experiment was the method used to implement the Bell basis measurement. In NMR, the measurement step allows us to measure the expectation values of σ_x and σ_y for each spin, averaged over the ensemble of molecules, rather than performing a projective measurement in some basis. For this reason, we must modify the projective measurement step in the standard description of teleportation, while preserving the teleportation effect.

We use a procedure inspired by Brassard *et al.*²⁰, who suggest a two-part procedure for performing the Bell basis measurement. Part one of the procedure is to rotate from the Bell basis into the computational basis, $|00\rangle$, $|01\rangle$, $|10\rangle$, $|11\rangle$. We implement this step in NMR by using the natural spin–spin coupling between the carbon nuclei, and r.f. pulses. Part two of the procedure is to perform a projective measurement in the computational basis. As Brassard *et al.* point out, the effect of this two-part procedure is equivalent to performing the Bell basis measurement, and leaving the data and ancilla qubits in one of the four states, $|00\rangle$, $|01\rangle$, $|10\rangle$, $|11\rangle$, corresponding to the different measurement results.

We cannot directly implement the second step in NMR. Instead, we exploit the natural phase decoherence occurring on the carbon nuclei to achieve the same effect. We note that phase decoherence completely randomizes the phase information in these nuclei and thus will destroy coherence between the elements of the above basis. Its effect on the state of the carbon nuclei is to diagonalize the state in the computational basis. In terms of the density matrix ρ for the carbon nuclei this process may be expressed as:

$$\begin{aligned} \rho \rightarrow & |00\rangle\langle 00|\rho|00\rangle\langle 00| + |01\rangle\langle 01|\rho|01\rangle\langle 01| + |10\rangle\langle 10|\rho|10\rangle\langle 10| \\ & + |11\rangle\langle 11|\rho|11\rangle\langle 11| \end{aligned} \quad (5)$$

As emphasized by Zurek²¹, the decoherence process is indistinguishable from a measurement in the computational basis for the carbons accomplished by the environment. We do not observe the result of this measurement explicitly, but the state of nuclei selected by the decoherence process contains the measurement result, and therefore we can do the final transformation conditional on the particular state the environment has selected. As in the scheme of Brassard *et al.*, the final state of the carbon nuclei is one of the four states, $|00\rangle$, $|01\rangle$, $|10\rangle$, $|11\rangle$, corresponding to the four possible results of the measurement.

Our experiment exploits the natural decoherence properties of the TCE molecule. The phase decoherence times (T_2) for C1 and C2 are approximately 0.4 s and 0.3 s. All other T_2 and T_1 times for all three nuclei are much longer, with a T_2 time for the hydrogen of ~ 3 s, and relaxation times (T_1) of approximately 20–30 s for the carbons, and 5 s for the hydrogen.

Thus, for delays of the order of 1 s, we can approximate the total evolution by exact phase decoherence on the carbon nuclei. The total scheme therefore implements a measurement in the Bell basis, with the result of the measurement stored as classical data on the carbon nuclei following the measurement. We can thus teleport the state from the carbon to the hydrogen and verify that the final state decays at the hydrogen rate and not the carbon rate.

Re-examining Fig. 1a we see how remarkable teleportation is from this point of view. During the stage labelled “measure in the Bell basis” in Fig. 1a, we allow the C1 and C2 nuclei to decohere and thus be measured by the environment, destroying all phase information on the data and ancilla qubits. Experimentally, a standard NMR technique known as refocusing employs r.f. pulses to ensure that the data qubit effectively does not interact with the target qubit. Classical intuition therefore tells us that the phase information about the input state, $|\Psi\rangle$, has been lost forever. Nevertheless, quantum mechanics predicts that we are still able

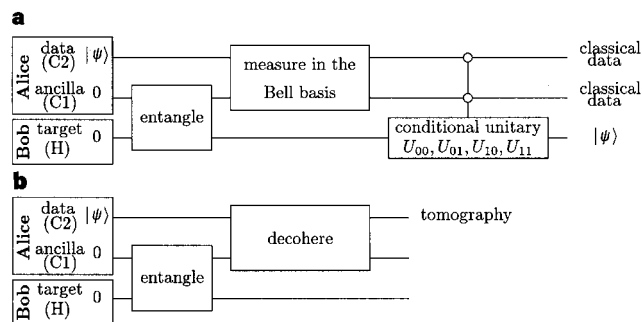


Figure 1 Schematic circuits for the quantum teleportation experiment (a) and the control experiment (b). The teleportation circuit is based on that suggested by Brassard *et al.*²⁰ We note that the control circuit omits two elements of the teleportation experiment—rotation from the Bell basis into the computational basis, immediately before the decoherence step, and the conditional unitary operation. Pulse sequences for our experiment are available at <http://www.cas.phys.unm.edu/~mnielsen/nmr/index.html>

to recover the complete system after this decoherence step, by quantum teleportation.

We implemented this scheme in TCE using a Bruker DRX-500 NMR spectrometer. Experimentally, we determined the Larmor and coupling frequencies for the hydrogen, C1 and C2 to be:

$$\omega_H \approx 500.133491 \text{ MHz}; \quad \omega_{C1} \approx 125.772580 \text{ MHz}; \quad (6)$$

$$\omega_{C2} \approx \omega_{C1} - 911 \text{ Hz}$$

$$J_{HC1} \approx 201 \text{ Hz}; \quad J_{C1C2} \approx 103 \text{ Hz} \quad (7)$$

The coupling frequencies between H and C2, as well as the chlorines to H, C1 and C2, are much lower, of the order of 10 Hz for the former, and less than 1 Hz for the latter. Experimentally, these couplings are suppressed by multiple refocusing, and will be ignored in the sequel. We note that C1 and C2 have slightly different frequencies, due to the different chemical environments of the two atoms.

We performed two separate sets of experiments. In one set, the full teleportation process was executed, making use of a variety of decoherence delays in place of the measurement. The readout was performed on the hydrogen nucleus, and a figure of merit—the entanglement fidelity—was calculated for the teleportation process. The entanglement fidelity is a quantity in the range 0–1 which measures the combined strength of all noise processes occurring during the process^{22,23}. (In the notation of ref. 22, we calculate $F_e(I/2, \varepsilon)$, where ε is the teleportation operation.) In particular, an entanglement fidelity of 1 indicates perfect teleportation, while an entanglement fidelity of 0.25 indicates total randomization of the state. Perfect 'classical transmission' corresponds to an entanglement fidelity of 0.5 (refs 22, 23), so entanglement fidelities greater than 0.5 indicate that teleportation of some quantum information is taking place.

The second set of experiments was a control set (Fig. 1b). In these experiments, only the state preparation and initial entanglement of H and C1 were performed, followed by a delay for decoherence on C1 and C2. The readout was performed in this instance on C2, and the entanglement fidelity was calculated for the process.

The results of our experiment are shown in Fig. 2, where the entanglement fidelity is plotted against the decoherence delay. Errors in our experiment arise from the strong coupling effect, imperfect calibration of r.f. pulses, and r.f. field inhomogeneities.

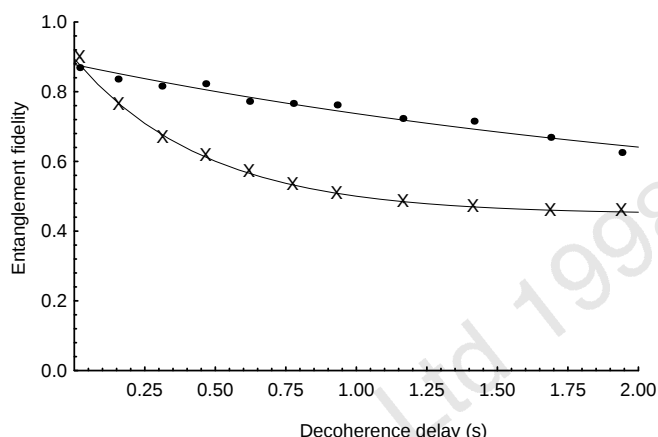


Figure 2 Entanglement fidelity (a measure of how well quantum information is preserved) plotted as a function of decoherence delay. The bottom curve is a control run where the information remains in C2; the curve shows a decay time of ~ 0.5 s. The top curve represents the fidelity of the quantum teleportation process. The decay time is ~ 2.6 s. The information is preserved for a longer time, corresponding approximately to the combined effects of decoherence and relaxation for the hydrogen, confirming the prediction of teleportation.

The estimated uncertainties in the entanglement fidelities are less than ± 0.05 , and are due primarily to r.f. field inhomogeneity and imperfect calibration of r.f. pulses.

To determine the entanglement fidelities for the teleportation and control experiments, we performed 'quantum process tomography' (refs 24, 25), a procedure for obtaining a complete description of the dynamics of a quantum system, as follows: the linearity of quantum mechanics implies that the single-qubit input and output for the teleportation process are related by a linear quantum operation²⁶. By preparing a complete set of four linearly independent initial states, and measuring the corresponding states output from the experiment, we may completely characterize the quantum process, enabling us to calculate the entanglement fidelity for the process²⁵.

We note three elements in Fig. 2. First, for small decoherence delays, the entanglement fidelity for the teleportation experiments significantly exceeds the value of 0.5 for perfect classical transmission of data, indicating successful teleportation of quantum information from C2 to H, with reasonable fidelity. Second, the entanglement fidelity decays very quickly for the control experiments as the delay is increased. Theoretically, we expect this to be the case, due to a T_2 time for C2 of ~ 0.3 s. Third, the decay of the entanglement fidelity for the teleportation experiments occurs much more slowly. Theoretically, we expect this decay to be due mainly to the effect of phase decoherence and relaxation on the hydrogen. Our experimental observations are consistent with this prediction, and provide more support for the claim that quantum data are being teleported in these experiments.

Note added in proof: Since completing this work we have become aware of related work by Furusawa *et al.*²⁷. □

Received 7 July; accepted 15 October 1998.

- Bennett, C. H. Quantum information and computation. *Phys. Today* **48**, 24–30 (1995).
- Preskill, J. Quantum computing: pro and con. *Proc. R. Soc. Lond. A* **454**, 469–486 (1998).
- Bennett, C. H. *et al.* Teleporting an unknown quantum state via dual classical and EPR channels. *Phys. Rev. Lett.* **70**, 1895–1899 (1993).
- Bouwmeester, D. *et al.* Experimental quantum teleportation. *Nature* **390**, 575–579 (1997).
- Boschi, D., Branca, S., De Martini, F., Hardy, L. & Popescu, S. Experimental realization of teleporting an unknown pure quantum state via dual classical and Einstein-Podolski-Rosen channels. *Phys. Rev. Lett.* **80**, 1121–1125 (1998).
- Sakurai, J. J. *Modern Quantum Mechanics* (Addison-Wesley, Reading, MA, 1995).
- Bennett, C. H., DiVincenzo, D. P., Smolin, J. A. & Wootters, W. K. Mixed state entanglement and quantum error correction. *Phys. Rev. A* **54**, 3824–3851 (1996).
- Cirac, J. I., Zoller, P., Kimble, H. J. & Mabuchi, H. Quantum state transfer and entanglement distribution among distant nodes in a quantum network. *Phys. Rev. Lett.* **78**, 3221–3224 (1997).
- Cory, D. G., Fahmy, A. F. & Havel, T. F. Ensemble quantum computing by NMR spectroscopy. *Proc. Natl Acad. Sci. USA* **94**, 1634–1639 (1997).

10. Ernst, R., Bodenhausen, G. & Wokaun, A. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions* (Oxford Univ. Press, 1990).
11. Grant, D. M. & Harris, R. K. (eds) *Encyclopedia of Nuclear Magnetic Resonance* (Wiley, New York, 1996).
12. Gershenfeld, N. & Chuang, I. L. Bulk spin resonance quantum computation. *Science* **275**, 350–356 (1997).
13. Cory, D. G., Price, M. D. & Havel, T. F. Nuclear magnetic resonance spectroscopy: An experimentally accessible paradigm for quantum computing. *Physica D* **120**, 82–101 (1998).
14. Laflamme, R., Knill, E., Zurek, W. H., Catagati, P. & Mariappan, S. V. S. NMR Greenberger-Horne-Zeilinger states. *Phil. Trans. R. Soc. Lond. A* **356**, 1941–1947 (1998).
15. Cory, D. G. *et al.* Experimental quantum error correction. *Phys. Rev. Lett.* **81**, 2152–2155 (1998).
16. Chuang, I. L., Gershenfeld, N. & Kubinec, M. Experimental implementation of fast quantum searching. *Phys. Rev. Lett.* **18**, 3408–3411 (1998).
17. Jones, J. A. & Mosca, M. Implementation of a quantum algorithm on a nuclear magnetic resonance quantum computer. *J. Chem. Phys.* **109**, 1648–1653 (1998).
18. Chuang, I. L., Vandersypen, L. M. K., Zhou, X. L., Leung, D. W. & Lloyd, S. Experimental realization of a quantum algorithm. *Nature* **393**, 143–146 (1998).
19. Jones, J. A., Mosca, M. & Hansen, R. H. Implementation of a quantum search algorithm on a nuclear magnetic resonance quantum computer. *Nature* **393**, 344–346 (1998).
20. Brassard, G., Braunstein, S. & Cleve, R. Teleportation as a quantum computation. *Physica D* **120**, 43–47 (1998).
21. Zurek, W. H. Decoherence and the transition from quantum to classical. *Phys. Today* **44**, 36–44 (1991).
22. Schumacher, B. W. Sending entanglement through noisy quantum channels. *Phys. Rev. A* **54**, 2614–2628 (1996).
23. Barnum, H., Nielsen, M. A. & Schumacher, B. W. Information transmission through a noisy quantum channel. *Phys. Res. A* **57**, 4153–4175 (1998).
24. Poyatos, J. F., Cirac, J. I. & Zoller, P. Complete characterization of a quantum process: the two-bit quantum gate. *Phys. Rev. Lett.* **78**, 390–393 (1997).
25. Chuang, I. L. & Nielsen, M. A. Prescription for experimental determination of the dynamics of a quantum black box. *J. Mod. Opt.* **44**, 2455–2467 (1997).
26. Nielsen, M. A. & Caves, C. M. Reversible quantum operations and their application to teleportation. *Phys. Rev. A* **55**, 2547–2556 (1997).
27. Furusawa, A. *et al.* Unconditional quantum teleportation. *Science* (in the press).

Acknowledgements. We thank D. Cory, C. Jarzynski, J. Ye and W. Zurek for discussions, the Stable Isotope Laboratory at Los Alamos for use of their facility, and the National Security Agency and Office of Naval Research for support.

Correspondence and requests for materials should be addressed to M.A.N. (e-mail: mnielsen@theory.caltech.edu).

Experimental verification of the quasi-unit-cell model of quasicrystal structure

Paul J. Steinhardt*, H.-C. Jeong†, K. Saitoh‡, M. Tanaka‡, E. Abe§ & A. P. Tsai§

* Department of Physics, Princeton University, Princeton, New Jersey 08544, USA

† Department of Physics, Sejong University, Kwangjin, Seoul 143-747, Korea

‡ Research Institute for Scientific Measurements, Tohoku University,

2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan

§ National Research Institute for Metals, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan

The atomic structure of quasicrystals¹—solids with long-range order, but non-periodic atomic lattice structure—is often described as the three-dimensional generalization of the planar two-tile Penrose pattern². Recently, an alternative model has been proposed^{3–5} that describes such structures in terms of a single repeating unit^{3–5}—the three-dimensional generalization of a pattern composed of identical decagons. This model is similar in concept to the unit-cell description of periodic crystals, with the decagon playing the role of a ‘quasi-unit cell’. But, unlike the unit cells in periodic crystals, these quasi-unit cells overlap their neighbours, in the sense that they share atoms. Nevertheless, the basic concept of unit cells in both periodic crystals and quasicrystals is essentially the same: solving the entire atomic structure of the solid reduces to determining the distribution of atoms in the unit cell. Here we report experimental evidence for the quasi-unit-cell model by solving the structure of the decagonal quasicrystal $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$. The resulting structure is consistent with images obtained by electron and X-ray diffraction, and agrees with the measured stoichiometry, density and symmetry of the compound. The quasi-unit-cell model provides a significantly better fit to these results than all previous alternative models, including Penrose tiling.

The fact that the entire structure reduces to a single repeating unit means that quasicrystals have a simplicity more like that of crystals than previously recognized. The quasi-unit-cell picture also suggests reasons why quasicrystals may form instead of crystals under some conditions. For the two-dimensional analogue, a pattern composed of overlapping decagonal quasi-unit cells, Gummelt³ and Steinhardt and Jeong^{4,5} have proved that the quasicrystal pattern is uniquely forced if decagons only overlay according to the rules shown in Fig. 1a. Steinhardt and Jeong^{4,5} have further proved that the quasi-unit-cell structure can arise simply by maximizing the density of a chosen cluster of atoms. If one imagines that the atomic cluster is energetically preferred, then the quasicrystal could be the ground-state configuration as minimizing the free energy maximizes the cluster density. Establishing the correspondence between quasi-unit cells and real atomic structures opens the door to theoretical and empirical tests of these notions.

The quasi-unit-cell and Penrose-tile pictures are examples of real-space descriptions of quasicrystals in which the structure is defined by decorating each quasi-unit cell or each type of tile identically. Some Penrose-tile models are based on the two-tile rhombus or kite-and-dart sets; others are based on the six-tile pentagonal set. In the hyperspace description, the quasicrystal is viewed as a projection from a higher-dimensional periodic, hypercubic lattice and the atomic decoration is defined in terms of atomic surfaces in the higher-dimensional space (five or six dimensions) which project into point atoms in three dimensions⁶. The surfaces may be continuous, discontinuous, or even fractal. Depending on the surfaces, the surfaces may produce a structure equivalent to decorating all tiles (or quasi-unit cells) identically or to decorating tiles differently depending on their local environment. Although the descriptions are mathematically related, the quasi-unit-cell picture is simpler in practice: it is much easier to solve for the structure by considering atomic rearrangements within a single quasi-unit cell in real, three-dimensional space than by considering simultaneously decorations or two or more tiles or by imagining surfaces in five or six dimensions.

To test the quasi-unit-cell hypothesis, we consider $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$, one of the best-characterized quasicrystal materials. $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ is an example of a decagonal phase; that is, a structure consisting of a periodic stack of ten-fold symmetric quasiperiodically ordered layers. Viewed along the periodic axis, the solid has the same symmetry as a Penrose tiling or an overlapping decagon structure. Hence it can be modelled as a columnar stacking of prisms with the shapes of Penrose tiles or decagons. Although quasicrystalline Al–Ni–Co has been studied for nearly a decade, deciphering its structure has been challenging for materials science reasons. The quasicrystal phase was originally found in a wide compositional range⁷. High-resolution transmission electron microscopy revealed decagonal cluster columns roughly 2 nm across⁸. Models of the structure based on two Penrose rhombus tiles, the Penrose six-tile pentagonal set^{9,10} and random packing of decagons were proposed on the basis of this early data¹¹. In random packing, there are no rules which force the decagons into a unique structure, so infinitely many distinct configurations are possible. (Similar random cluster packing models have been developed for other quasicrystalline materials with decagonal and icosahedral symmetry^{12–14}.) Although the random cluster packing models and the quasi-unit-cell picture discussed here share the idea of building the atomic structure from a repeating unit, they differ in three important ways: first, the random decagon clusters do not cover the entire structure; second, the random decagon clusters are imperfectly ordered and their arrangement is not unique; and third (in most cases), the random decagon clusters contain an atomic configuration that is five- or ten-fold symmetric.

But it has recently become apparent that, over most of the compositional and temperature range, one obtains an inhomogeneous mixture of structures^{15–17}. The ideal, high-perfection phase

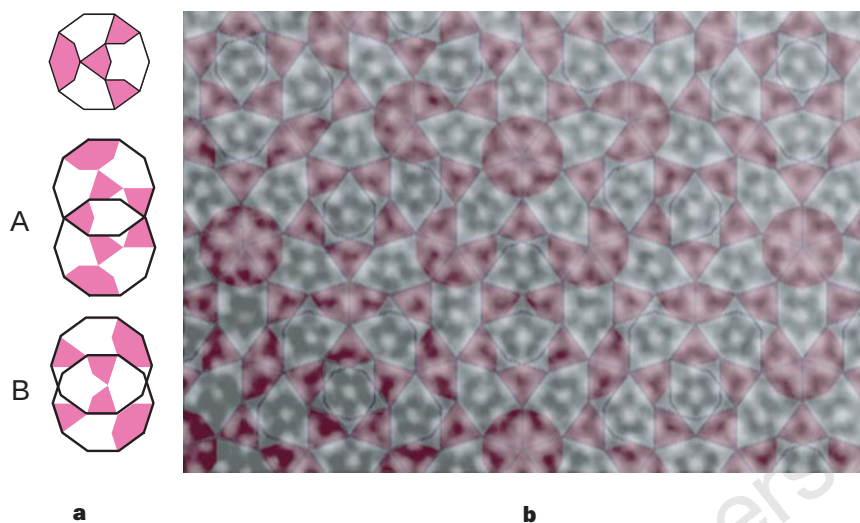


Figure 1 Comparison of atomic lattice image and quasi-unit-cell picture. **a**, The decagon quasi-unit cell with coloured kite-shapes indicating overlap rules (overlapping regions must have matching colour); only 'A' or 'B' overlaps are permitted. **b**, Two overlaying images. In black and white is the high-angle annular dark-field (HAADF) image of $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ in which decagonal rings of atoms are evident. To obtain an objective measure, we digitized the image and performed a random search for rings of ten atoms (white spots) of the type shown in Fig. 2. Allowing modest tolerance for distortion and digitization error, the random search

converged on a set of decagons that cover the entire image. The decagon covering has been superposed over the HAADF image in **b**. All decagons overlap neighbours according to A or B overlaps and the decagon tiling is a perfect quasiperiodic pattern. Furthermore, the decagon interiors have been coloured with the kite-shapes which serve as mnemonics for the overlap rules, as shown in **a**. In **b**, there is near-perfect correspondence between the atomic decorations of both the decagons and the kite-regions.

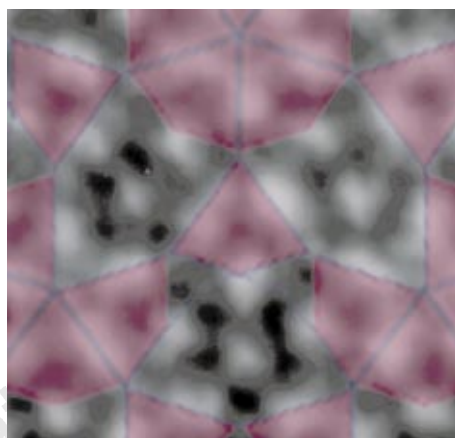


Figure 2 Magnified view of decagon cluster displaying broken symmetry. A magnified view of a decagonal cluster from Fig. 1 illustrating a decagonal ring of atoms (at edge-centres of decagon) and showing how the three bright spots in the middle of the decagonal cluster break decagonal symmetry in a way similar to the kite-representation of the overlap rules.

has a composition $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$, significantly different from the samples on which previous models were based. Whereas earlier samples exhibited diffuse scattering and superlattice reflections in the diffraction patterns, $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ does not. The phase is stable and reproducible at temperatures of about 800°C , and highly perfect water-quenched samples can be grown as large as 1 mm. So it is only now that precise measurements of the structure, stoichiometry and density of this material have become possible, which provide powerful new constraints on models. Previous models fail badly on all counts. (Saitoh has recently developed a model based on two kinds of pentagon clusters^{18,19}—the first to reproduce the observed breaking of ten-fold symmetry inside the 2-nm clusters and to obtain the correct space-group symmetry—but his model does not match the measured density or stoichiometry.)

To test the quasi-unit-cell picture, we have performed the following tests.

(1) *Decagon overlap test*. In the quasi-unit-cell picture, the structure

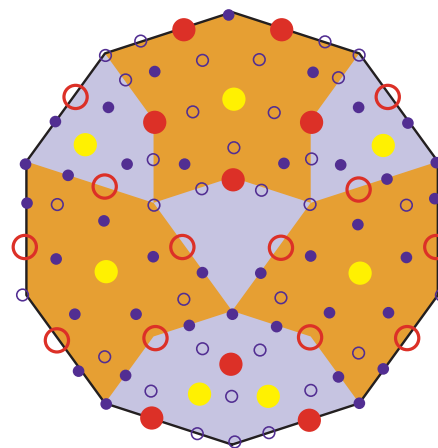


Figure 3 The best-fit candidate model for the atomic decoration of the decagonal quasi-unit cell for $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$. Large circles represent Ni (red) or Co (yellow) and small circles represent Al. The structure has two distinct layers along the periodic c -axis. Solid circles represent $c = 0$ and open circles represent $c = 1/2$. The experimental input used to generate candidate models is: (1) the space group $P10_5/mmc$ based on convergent-beam electron-diffraction methods¹⁸; (2) the HAADF image, in which scattering is proportional to the square of the atomic number, highlights the positions of the transition metals relative to the aluminium; and (3), based on X-ray diffraction, the diameter of the decagonal cluster is 2.033 nm and the period along the decagonal c -axis is 0.408 nm, corresponding to two atomic layers per period. To obtain $P10_5/mmc$, different atomic arrangements are placed on the two layers along the c -axis such that two overlapping decagons related by odd-multiple rotations of $\pi/10$ must be shifted by a half-period in order to share atoms. Thus the covering of the plane can be decomposed into two interpenetrating lattices of decagons, in which the decagon decoration in one set is related by screw symmetry or c -glide symmetry with respect to the other. Using positions of the transition metals inferred from the HAADF images, a sequence of plausible decagon decorations was explored and the best fits to the experimentally measured stoichiometry and density ($3.94 \text{ g cm}^{-3} \pm 1\%$) were found. For each decoration, the constraint had to be imposed that shared atoms at overlapping sites were identical.

consists of a covering by a single, repeating decagonal cluster (column) which overlaps (shares atoms with) neighbours in two, and only two, distinct ways ('A' and 'B'), according to the overlap rules⁴ shown in Fig. 1a. In Fig. 1b, we digitized the high-angle annular dark-field (HAADF) image¹⁸ of $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ and searched randomly for decagonal rings of atoms (white spots). The random search converged on a set of decagons that cover the entire image in a perfect quasiperiodic pattern in which all decagons overlap neighbours according to A or B overlaps.

(2) *Broken decagonal symmetry test.* In Fig. 1a, the decagon interior has been coloured with kite-shapes which break ten-fold symmetry. The kite-shapes serve as mnemonics for the overlap rule. Figure 2 is a magnified view of one decagon from Fig. 1b which shows that the central triangle of atoms breaks the ten-fold symmetry^{18,20} in a way that is isomorphic to the kite-shapes. The correspondence is maintained throughout Fig. 1b. In principle, one could have overlapping decagons with ten-fold (or five-fold) symmetric decorations; however, the broken symmetry is significant because it suggests that overlap rules may play a role in determining the atomic arrangements.

(3) *Stoichiometry/density/space-group tests.* For any candidate model of $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$, the atomic structure is totally determined by the atomic decoration of the decagonal quasi-unit cell. Steinhardt and Jeong (manuscript in preparation) have developed a procedure for computing the stoichiometry, density and symmetry given the atomic decoration of the quasi-unit cell. The procedure is non-trivial because neighbouring cells share atoms, and double-counting must be avoided. This procedure was used to compare models with the experimentally determined stoichiometry, density and symmetry. Figure 3 is, by a significant margin, the best-fit model. The density of 3.97 g cm^{-3} and the stoichiometry of $\text{Al}_{72.5}\text{Ni}_{18.6}\text{Co}_{8.9}$ are both within the reported experimental uncertainties and the space-group symmetry is correct. For this material, the best-fit model happens to be one which cannot be described as a simple decoration of a Penrose rhombus tiling in which each type of tile is

decorated identically. Decorations of Penrose tiles are a proper subset of the possible decorations of (decagon) quasi-unit cells (P.J.S. and H.-C.J., manuscript in preparation) and so they are automatically included in our analysis. No Penrose model satisfying the HAADF and symmetry constraints was found which matched the measured stoichiometry and density.

(4) *HREM image test.* We computed high-resolution electron microscopy (HREM) images for our best-fit model, and compared them with experimental images. Figure 4 shows that the agreement is excellent. This test is somewhat independent of the first test because HREM images include contributions of both aluminum and transition-metal atoms, whereas HAADF images display prominently only transition metals.

The quasi-unit-cell picture for the decagonal phase of $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ has passed all currently possible tests; it satisfies more constraints than previous atomic models for $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ (or other quasicrystals) and provides a significantly better fit than previous approaches. We believe that other decagonal quasicrystals will also fit well. For example, a new class of stable decagonal quasicrystals belonging to the Frank-Kasper family have been discovered, including ZnMgDy (ref. 21), which also produce images composed of overlapping decagonal clusters whose internal structure breaks decagonal symmetry in accordance with the overlap rules. The ultimate aim is to move beyond using the quasi-unit-cell picture to describe the structure, and to begin to explain why the structure forms in the first place. The quasi-unit-cell picture suggests that the key is understanding the energetics of the repeating atomic cluster: deciphering quasicrystals in terms of quasi-unit cells should therefore provide realistic data to test theories of why quasicrystals form. □

Received 8 July; accepted 19 August 1998.

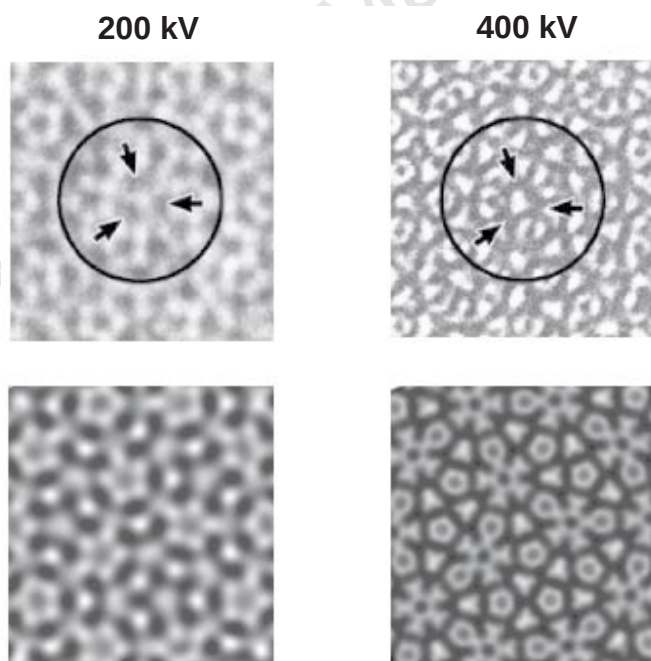


Figure 4 Comparison of experimental lattice images with computed image for best-fit model. A comparison of the high-resolution electron microscope images of $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ (top row) at accelerating voltages of 200 kV (obtained by Hiraga *et al.*⁸) and 400 kV (obtained by Saitoh *et al.*¹⁹), compared to computed images (bottom row) for the candidate model in Fig. 3. The images were computed using the software package MacTempas (Total Resolution Inc., Berkeley, CA). Allowing for minor imperfections in the experimental sample and our inability to match precisely the observational conditions, the correspondence is very close.

- Levine, D. & Steinhardt, P. J. Quasicrystals: a new class of ordered structures. *Phys. Rev. Lett.* **53**, 2477–2480 (1984).
- Penrose, R. The role of aesthetics in pure and applied mathematical research. *Bull. Inst. Math. Applic.* **10**, 266–271 (1974).
- Gummelt, P. Construction of Penrose tilings by a single aperiodic prototile. *Geometriae Dedicata* **62**, 1–17 (1996).
- Steinhardt, P. J. & Jeong, H.-C. A simpler approach to Penrose tiling with implications for quasicrystal formation. *Nature* **382**, 433–435 (1996).
- Jeong, H.-C. & Steinhardt, P. J. Constructing Penrose-like tilings from a single proto-tile and the implications for quasicrystals. *Phys. Rev. B* **55**, 3520–3532 (1997).
- Janot, C. & de Boissieu, M. in *Quasicrystals: The State of the Art* (eds DiVincenzo, D. & Steinhardt, P. J.) 57–94 (World Scientific, Singapore, 1995).
- Tsai, A. P., Inoue, A. & Masumoto, T. Icosahedral, decagonal and amorphous phases in Al–Cu–M (M = transition metal) systems. *Mater. Trans. Jpn. Inst. Metals* **30**, 666–676 (1989).
- Hiraga, K., Lincoln, E. J. & Sun, W. Structure and structural change of Al–Ni–Co decagonal quasicrystal by high-resolution electron microscopy. *Mater. Trans. Jpn. Inst. Metals* **32**, 308–314 (1991).
- Yamamoto, A. Structure of decagonal $\text{Al}_{65}\text{Cu}_{20}\text{Co}_{15}$ quasicrystals. *Sci. Rep. Res. Inst. Tohoku Univ.* **42**, 207–212 (1996).
- Ritsch, S. *et al.* Highly perfect decagonal Al–Co–Ni quasicrystals. *Phil. Mag. A* **74**, 99–106 (1996).
- Burkov, S. Modeling decagonal quasicrystals: random assembly of interpenetrating decagonal clusters. *J. Phys.* **2**, 695–706 (1992); Structure model of the Al–Cu–Co decagonal quasicrystal. *Phys. Rev. Lett.* **67**, 614–617 (1991).
- Steurer, W. & Kuo, K. H. Five-dimensional structure analysis of decagonal $\text{Al}_{65}\text{Cu}_{20}\text{Co}_{15}$. *Acta Crystallogr. B* **46**, 703–712 (1990).
- Boudard, M. *et al.* Atomic structure of the Al–Pd–Mn icosahedral phase. *J. Non-Cryst. Solids* **153–4**, 5–9 (1993).
- Janot, C. & Patera, J. Simple physical generation of quasicrystals. *Phys. Lett. A* **233**, 110–114 (1997).
- Edagawa, K., Ichihara, M., Suzuki, K. & Takeuchi, S. New type of decagonal quasicrystal with superlattice order in Al–Ni–Co alloy. *Phil. Mag. Lett.* **66**, 19–25 (1992).
- Ritsch, S., Beeli, C., Nissen, H.-U. & Luck, R. Two different superstructures of the decagonal Al–Co–Ni quasicrystal. *Phil. Mag. A* **71**, 671–685 (1995).
- Fujiwara, A., Tsai, A. P. & Inoue, A. in *Proc. 6th Int. Conf. Quasicrystals* (eds Takeuchi, S. & Fujiwara, A.) 341 (World Scientific, Singapore, 1988).
- Saitoh, K., Tsuda, K., Tanaka, M., Kaneko, K. & Tsai, A. P. Structural study of an $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ decagonal quasicrystal using the high-angle annular dark-field method. *Jpn. J. Appl. Phys.* **36**, L1400–L1402 (1997).
- Saitoh, K., Tsuda, K. & Tanaka, M. A new structural model of an $\text{Al}_{72}\text{Ni}_{20}\text{Co}_8$ decagonal quasicrystal. Preprint (Tohoku Univ., 1998); *J. Phys. Soc. Jpn.* (in the press).
- Beeli, C. & Horiuchi, S. The structure and its reconstruction in the decagonal $\text{Cl}_{70}\text{Cu}_{17}\text{Co}_{13}$. *Phil. Mag. B* **70**, 215–240 (1994).
- Sato, T. J., Abe, E. & Tsai, A. P. A novel decagonal quasicrystal in Zn–Mg–Dy system. *Jpn. J. Appl. Phys.* **36**, L1038–L1039 (1997).
- Mermin, N. D. in *Quasicrystals: The State of the Art* (eds DiVincenzo, D. & Steinhardt, P. J.) 133–184 (World Scientific, Singapore, 1991).

Acknowledgements. We thank R. Kilaas of Total Resolution, Inc. for lending us the MacTempas software package for this project. This work was partially supported by the US Department of Energy at Princeton, and by CREST (Core Research for Evolutional Science and Technology) of Japan Science and Technology Corporation (JST).

Correspondence and requests for materials should be addressed to P.J.S. (e-mail: steinh@princeton.edu).

Interface structure between silicon and its oxide by first-principles molecular dynamics

Alfredo Pasquarello^{*†‡}, Mark S. Hybertsen[‡] & Roberto Car^{*†}

^{*} Institut Romand de Recherche Numérique en Physique des Matériaux (IRRMA), Ecublens, CH-1015 Lausanne, Switzerland

[†] Department of Condensed Matter Physics, University of Geneva, CH-1211 Geneva, Switzerland

[‡] Bell Laboratories, Lucent Technologies, 700 Mountain Avenue, Murray Hill, New Jersey 07974, USA

The requirement for increasingly thin (<50 Å) insulating oxide layers in silicon-based electronic devices highlights the importance of characterizing the Si–SiO₂ interface structure at the atomic scale. Such a characterization relies to a large extent on an understanding of the atomic-scale mechanisms that govern the oxidation process. The widely used Deal–Grove model invokes a two-step process in which oxygen first diffuses through the amorphous oxide network before attacking the silicon substrate, resulting in the formation of new oxide at the buried interface¹. But it remains unclear how such a process can yield the observed near-perfect interface^{2–12}. Here we use first-principles molecular dynamics^{13–15} to generate a model interface structure by simulating the oxidation of three silicon layers. The resulting structure reveals an unexpected excess of silicon atoms at the interface, yet shows no bonding defects. Changes in the bonding network near the interface occur during the simulation via transient exchange events wherein oxygen atoms are momentarily bonded to three silicon atoms—this mechanism enables the interface to evolve without leaving dangling bonds.

The atomic-scale structure of the interface between crystalline Si and its amorphous oxide remains controversial^{8–12}. This is a problem in view of the high degree of perfection achieved in practice, as evidenced by the very low density of defects observed at this interface (less than one per 10⁴ interface atoms)¹⁶. One of the puzzles which any structural model for this interface must resolve is the required factor of two reduction in silicon density upon passing from the crystalline silicon to the oxide. This occurs in a very short distance because by most measurements the interface is quite abrupt. The puzzle can be visualized by noting that a bulk (001) terminated silicon crystal would expose two dangling bonds per silicon atom, while the oxide at normal volume can easily accommodate only one bond.

The process leading to a disordered silica network is necessarily more involved than just the incorporation of oxygen atoms in Si–Si bonds of the silicon substrate. Such a process would lead to an unrealistic oxide with the β-cristobalite structure occupying less than half of its equilibrium volume. Thus, the high pressures occurring during the oxidation process are expected to be relieved by structural rearrangements causing bond breakings, which modify the ring statistics and ultimately give rise to the amorphous nature of the oxide. In fact for growth above about 950 °C, the oxide film grows stress-free, attributed to “viscous flow” of the oxide¹⁷. Experimental results obtained by medium-energy ion scattering^{5,6} also provide evidence for such rearrangements. These measurements show that oxygen atoms incorporated into the silica network move during subsequent oxidation, and are incorporated into the newly formed oxide. Such dynamical processes are missing from the Deal–Grove model.

A full atomistic simulation of oxidation at realistic temperatures and for realistic timescales is far beyond present capabilities. Here

we apply molecular dynamics at sufficiently high temperatures to explore low-energy atomic excitations of the structure. In this way, dynamical processes at Si(001)–SiO₂ interfaces leading to structural rearrangement are modelled. Starting from an idealized interface between Si and a crystalline form of SiO₂, the computer simulation generates a more realistic interface between Si and a disordered oxide. We use first-principles molecular dynamics^{13–15}, in which the electronic structure evolves self-consistently during the atomic motion. This is essential in order to describe the varying oxidation states of Si (refs 18–20) near the interface. Application of this method to bulk silica showed that high temperature only induces occasional disruptions of the network, and the simulation led to a structural configuration which compared well with experiment^{21,22}.

The initial simulation structure was obtained by attaching tridymite, a crystalline form of SiO₂, to a Si(001) substrate^{10–12}. For simulation, a thin-film system was defined, consisting of six layers of silicon (8 Å) and nine monolayers of oxide (11 Å) with a periodic interface arrangement (repeat distance 10.8 Å). The technical aspects concerning the electronic structure were the same as in refs 10–12. In order to preserve the crystalline structure of the substrate, the three lowest Si layers were coupled to a thermostat^{23,24} which was held at a fixed low temperature (1,000 K). The temperature of a 5-Å zone of oxide starting ~2 Å above the initial interface was controlled by a second thermostat. By increasing the temperature gradient, we could drive substantial atomic motion near the interface. In particular, oxygen atoms diffused out of the heated zone towards the interface, thereby exposing the Si substrate to a flux of oxygen atoms. However, no external oxygen atoms were supplied to further sustain oxidation.

For this type of complex system, only relatively short timescales of the order of a few tens of picoseconds can be simulated with present computers. We increased the temperature of the driving thermostat, looking for substantial atomic rearrangement on the picosecond timescale. When the driving thermostat reached a temperature sufficiently high to disrupt the network and to induce diffusive events (6,000 K), this thermal regime was maintained until the oxide structure departed from its initial crystalline topology (3 ps). Under these conditions the temperature in the neighbourhood of the interface ranged between 2,500 and 3,000 K. This allowed oxygen atoms to move into the Si substrate, resulting in the oxidation of three monolayers of silicon. Then, after a gradual quench, a relaxed model structure of the interface was obtained. Overall, the simulation lasted for 24 ps.

Although our simulation procedure necessarily involves high temperatures and short times, the final resulting structure is unbiased by preconceived ideas of the interface bonding. Our final quenched structure (Fig. 1) shows a smooth connection of amorphous oxide to crystalline silicon. There are no coordination defects in the neighbourhood of the Si substrate. The stoichiometry of the first 5 Å of oxide is close to SiO. The intermediate oxidation states of Si occur in roughly comparable amounts, in agreement with the distribution found in photoemission spectra^{18,19}. However, the thickness of the SiO layer and the total amount of the intermediate oxidation states of Si are a consequence of the conditions in the simulation, which did not allow for a sustained flux of oxygen atoms.

Our final quenched structure has surprising features that were not previously considered in models of the interface structure. The two lowest silicon layers (labelled in Fig. 1) show the same bond pattern as bulk Si(001). However, we find that the bond pattern in the third silicon layer is significantly altered by the presence of one-quarter of a monolayer of additional silicon atoms. An example where an extra Si atom has been incorporated into the lattice is highlighted by the unlabelled arrow in Fig. 1. At first glance, this is unexpected because an increase in silicon density would be naively expected to lead to an increase in bonds to be satisfied. However,

this effect is countered by the formation of Si–Si intralayer bonds, yielding in this way a net reduction in the number of bonds into the oxide. This is seen as bonds lying in the plane of the interface in Fig. 1. Furthermore, this dense Si layer is spread out over a thickness of about 1 Å along the *z* direction, facilitating the match with the disordered structure of the oxide. Such a silicon density increase at the interface is well within the limits set by ion-scattering experiments^{25,26} and is in qualitative agreement with recent X-ray reflectivity measurements²⁷. We note that the disordered bond pattern in the third silicon layer occurs without any oxygen having been incorporated into this layer.

An important *a posteriori* justification for using high temperatures, large temperature gradients and short time spans is provided by the bonding properties of the final interface structure, which compare well with those found in a previous theoretical study¹². In particular, the degree of disorder in the third Si layer appears reasonable, showing an average Si–Si bond length of $d = 2.34$ Å with a standard deviation of $\sigma = 0.082$ Å, in good accord with experimental results for amorphous silicon ($d = 2.36$ Å and $\sigma = 0.074$ Å)²⁸.

The increased density of Si atoms in the layer adjacent to the interface was a consequence of the detailed atomic motions that accompanied the movement of oxygen into the layer above. Despite the rather high temperatures, oxygen atoms formed bonds with two silicon atoms during most of the simulation time. A modification of the network topology occurs through exchange events by which one Si neighbour is replaced by another Si atom. This occurred in our simulation through an associative mechanism in which the oxygen atom was momentarily bonded to three Si atoms. The threefold-coordinated oxygen atom and its evolution are illustrated in Fig. 2. As shown in this example, these events often survived for several Si–O bond oscillations. The local atomic configuration is almost planar, with Si–O bonds of 1.8 Å and Si–O–Si bond angles of 120°. Threefold-coordinated oxygen events were accompanied by substantial rearrangements of the Si lattice, which often occurred through creation of either threefold or fivefold-coordinated Si atoms. Such defective structures correspond to low-energy structural excitations of amorphous silicon²⁹. Our simulation shows that excited configurations involving threefold-coordinated oxygen centres and either threefold- or fivefold-coordinated Si atoms produce a structural relaxation without leaving dangling bonds. The observed atomic motion cannot adequately be pictured by local jump-like processes involving single atoms, but should rather be described by group displacements of several strongly connected atoms. The net effect of oxygen penetration through a cascade of

events involving threefold-coordinated oxygen atoms was to expel Si atoms out of the newly oxidized Si layer. This occurred in our simulation either by driving new Si–Si bonds up into the oxide or by pushing Si atoms into the Si layer underneath the oxide. The latter mechanism yields a densified Si layer at the interface, and could relate to the emission of Si interstitials which is known to occur on oxidation⁷.

The timescales in our simulation are highly artificial because of the imposed temperature regime. However, owing to the realistic description of the potential-energy surface, we expect the sequence of atomic rearrangements to be meaningful. In particular, the network motion centred on the events involving threefold-coordinated oxygen which we have observed here could be representative of the dynamical processes near the interface. We now outline a scheme which illustrates the role of these dynamical processes which were previously missing from models of oxidation.

Deferring the issue of longer-range oxygen transport through the film, which would require substantial further work, we focus on dynamical rearrangements of the lattice. Incoming oxygen atoms

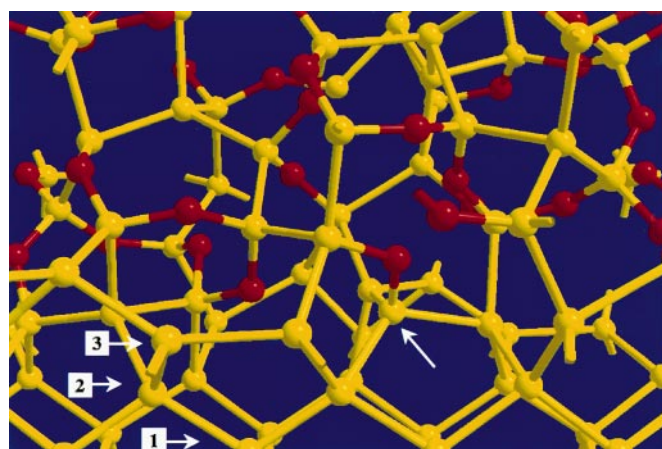


Figure 1 'Ball and stick' model of the Si(001)–SiO₂ interface structure after the final quench. Yellow and red balls indicate Si and O atoms, respectively.

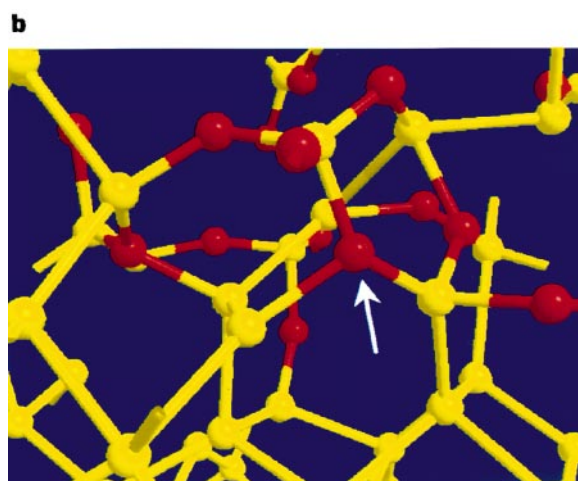
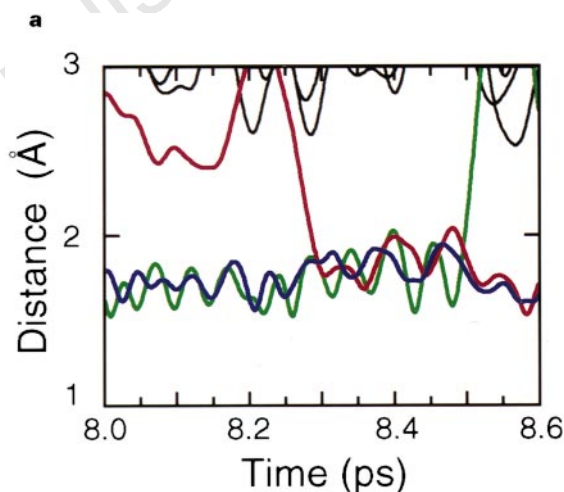


Figure 2 Threefold-coordinated O atom. **a**, The trajectories of Si atoms with respect to a given oxygen atom are shown as a function of time, illustrating an exchange event in the oxygen first-nearest-neighbour shell. The trajectories of different nearest neighbour Si atoms are shown by different colours. **b**, 'Ball and stick' model of the corresponding threefold-coordinated oxygen atom. Colours: same as in Fig. 1.

Received 7 May; accepted 17 August 1998.

- Acknowledgements.** We thank A. Mangili for providing visualization support. The calculations were performed on the NEC-SX4 of the Swiss Center for Scientific Computing (CSCS).

60

Figure 1 Structures of molecules **1–5**.

molecule is weakly coupled electronically to the donor. The fluorescence quantum yield of TET itself is 0.16 (ref. 22), yet the fluorescence of TET in **1–5** is strongly quenched, when each molecule is dissolved in 2-methyltetrahydrofuran (MTHF). This result indicates that the excited state of TET is being deactivated by a new mechanism in **1–5**. Reference molecules that lack the PI acceptor do not quench the lowest excited singlet state of TET.

Electron transfer within these compounds was investigated using femtosecond optical pump-probe spectroscopy²³. A pump laser pulse (wavelength 485 nm, length 150 fs) was used to selectively excite TET within **1–5** to its lowest excited singlet state. Transient spectra were measured with an overall 180-fs instrument response using a white-light continuum probe pulse²³. These spectra all develop a broad peak at 610 nm and sharper peaks at 720 and 870 nm as a function of time (Fig. 2b). The 720-nm peak is assigned to PI[−] on the basis of previous spectroelectrochemical measurements²⁴, while the 610-nm and 870-nm features are characteristic of the TET radical cation²⁵. Thus, the fluorescence quenching observed in these electron transfer systems is a consequence of electron transfer from ¹TET to PI. The measured charge separation (CS) and charge recombination (CR) rate constants for **1–5** are given in Table 1. All of the formation dynamics were fitted to an instrument-limited rise time due to the initial formation of ¹TET, and an exponential rise due to CS. The CR dynamics for **1–4** were fitted using a single exponential decay, while for **5** a reasonable fit to the data required a sum of two exponentials (Table 1).

The values of R_{DA} listed in Table 1 were obtained from structures optimized using the MM2 force field. In these structures the dihedral angle between TET and the phenyl group connecting it to the bridge molecules is 69°, while that between the bridge phenyl and PI is 54°. The dependence of the CS rate constants on increasing bridge length is given in Fig. 3a, where $\ln k_{CS}$ versus R_{DA} is shown for **1–5** in MTHF. There appear to be two electron-transfer regimes in these molecules. For the two shortest bridges, in **1** and **2**, there is a

Table 1 Data for charge separation and recombination reactions in **1–5**

Molecule	R_{DA} (Å ^{−1})	$-\Delta G_{CS}$ (eV)	$-\Delta G_{CR}$ (eV)	k_{CS} (s ^{−1})	k_{CR} (s ^{−1})
1	11.1	0.84	1.71	1.31×10^{11}	6.21×10^{10}
2	17.7	0.77	1.78	2.27×10^{10}	1.39×10^{10}
3	24.3	0.74	1.81	3.88×10^{11}	1.59×10^{10}
4	30.9	0.72	1.83	2.63×10^{11}	6.37×10^9
5	38.0	0.70	1.85	2.18×10^{11}	1.81×10^{10} (0.65) 5.81×10^9 (0.35)

The free energies of reaction are obtained from the one-electron redox potentials for oxidation of TET (0.77 vs SCE), reduction of PI (−0.79 V vs SCE), R_{DA} , and the lowest excited singlet state energy of TET (2.55 eV) using procedures given in ref. 23. For **5** the relative amplitudes of the two components of k_{CR} are given in parentheses.

sharp decrease in electron-transfer rate constant with distance, which is consistent with previous experimental results for other conjugated bridges^{16,18–21}. There is an abrupt change in mechanism starting with **3**, where the CS rate constant for **3** is greater than that for **1** even though the donor–acceptor distance of **3** is 13 Å longer than that of **1**. A least-squares fit for the three longest bridges yields a β value of only 0.04 Å^{−1}. The corresponding distance dependence of thermal CR is shown in Fig. 3b. In addition to the distance dependence of the electronic coupling between the donor and the acceptor, the free energy of reaction (ΔG) and the solvent reorganization energy (λ_s) are also distance dependent primarily as a consequence of the coulombic interaction between the ions. However, when donor–acceptor distances are significantly greater than 10 Å, as they are in **2–5**, the contribution of the distance dependence of ΔG and λ_s to the overall distance dependence of the electron-transfer rate constant is small²⁶. The CR rate constants for **2–5** decrease monotonically with distance, if the slower rate constant from the bi-exponential data for **5** is used. However, a break in this trend is observed if the faster rate constant for CR in **5** is considered.

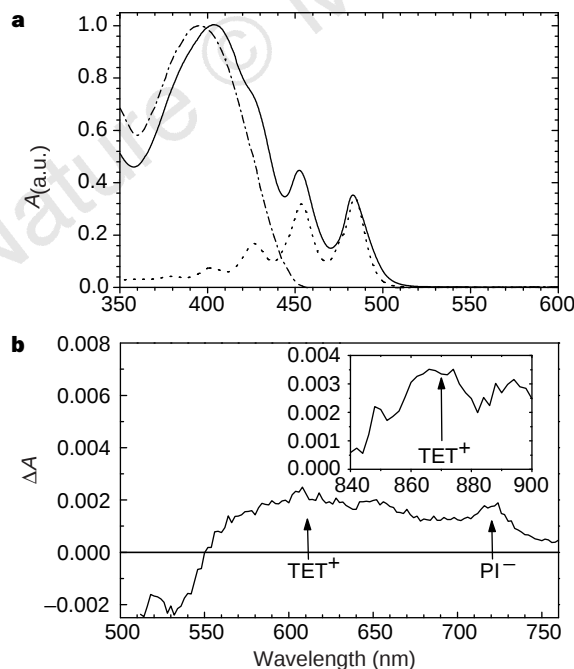


Figure 2 Optical absorption spectra. **a**, Ground-state absorption spectra of **3** (solid line), TET (dashed line), and PPV3-PI (dot-dashed line) in MTHF. **b**, Transient absorption spectrum of **3** in MTHF taken 150 ps after a 485 nm, 150 fs, 0.5-μJ pulse. The sample was contained in a cell of 2 mm path length within which the pump and probe beams overlapped in a 200-μm spot. Inset, near-infrared region of the spectrum at the same time delay. A , Absorbance of the sample.

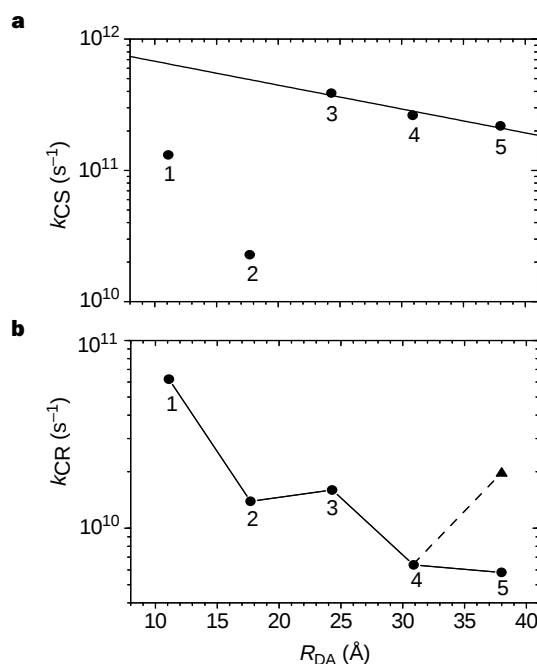


Figure 3 Distance dependence of charge separation and recombination rate constants. **a**, Plot of $\ln k_{CS}$ versus the donor–acceptor distance, R_{DA} in MTHF (filled circles). A Levenberg–Marquardt nonlinear least-squares fit to the data for molecules **3–5** is also shown. **b**, Plot of $\ln k_{CR}$ versus R_{DA} in MTHF (filled circles and filled triangles). Lines are used to connect the points.

The data suggest that CR in **5** proceeds by two different mechanisms (see below).

The importance to the efficiency of charge injection into the wire of nearly matching the energy of the highest occupied electronic state of the donor to the lowest unoccupied state of the molecular wire has long been recognized^{27–29}. Yet, experimental demonstrations of this phenomenon have been lacking. As the energy gap between the highest occupied state of the donor (usually thought of as the Fermi level of a metallic contact) and an empty eigenstate of the bridge decreases, the rate of charge conduction through the wire increases rapidly. We may apply a similar analysis to the DBA molecules studied here by examining the changes in the CS rates as the relevant injection energy gap between the donor and PPV bridge is modified by the length and electronic substitution pattern along the oligomeric PPV backbone. At a simple level of analysis, the electron transfer event can be envisioned as the movement of an electron from the lowest unoccupied molecular orbital (LUMO) of the photoexcited donor to the LUMO of the acceptor, via the LUMO of the bridge. Figure 4 is a plot of the bridge highest occupied molecular orbital (HOMO) and LUMO energies as a function of backbone length, as well as those of TET. In **1** and **2** the LUMO–LUMO energy gap of the donor and bridge is substantially larger than that in the three longer bridges. Lengthening the bridge in **3**, **4** and **5** causes the LUMO energy of the bridge to approach that of TET to within 0.1 eV. In addition to the bridge-length dependence, the appended alkoxy groups on the oligomeric backbone also change the HOMO and LUMO energies³⁰.

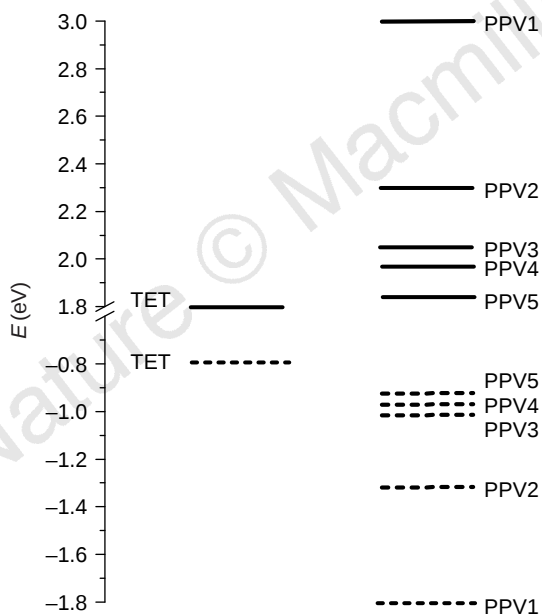
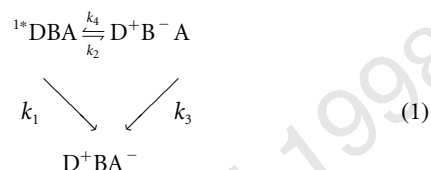


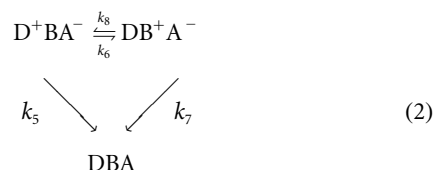
Figure 4 The LUMO (solid lines) and HOMO (dashed lines) energies of TET and the five bridge molecules. These energies were derived from the experimental spectroscopic and electrochemical properties of the individual chromophores. For the PPV3, PPV4 and PPV5 bridges, all-*trans* hydrogen-terminated bridge molecules were synthesized. The oxidation potentials of TET, PPV1 (benzene), PPV2 (stilbene), PPV3, PPV4 and PPV5 in butyronitrile with 0.1 M tetra-*n*-butylammonium perchlorate are 0.77, 1.80, 1.32, 1.01, 0.96 and 0.94 V versus SCE, respectively, while their lowest excited singlet state energies are 2.55, 4.82, 3.60, 3.01, 2.90 and 2.75 eV, respectively, as measured by techniques given in ref. 23. The sum of the oxidation and reduction potentials for each bridge should be approximately the same as the lowest excited singlet state energy of the bridge. Using this relationship, we equate the HOMO energy of a bridge molecule to the negative of its oxidation potential, and calculate its LUMO energy from the difference between its lowest excited singlet state energy and its oxidation potential.

The changes in the donor-to-bridge injection energy are reflected very well in the measured CS kinetics. In **1** and **2**, the energy gap between the donor and bridge LUMOs is large enough that the electron-transfer event exhibits the normal superexchange mechanism of electron transfer; that is, $k_1 \gg k_2$ in equation (1).



In the three longest bridges, this energy gap becomes small enough that there is very little energy penalty paid for injection of an electron from the donor into the bridge, and the rates decrease very slowly with distance. In **5** the energy gap is so small that the two LUMOs are almost isoenergetic. While this might be expected to lead to very strong electronic mixing between TET and the bridge in **5**, the 69° dihedral angle between ET and the first phenyl group of the bridge strongly diminishes electronic coupling between TET and the bridges, as is evidenced by the invariance of the lowest excited singlet state energy of TET in the series **1–5**. The very small energy gap suggests that incoherent electron transfer, in which the electron density is temporarily dephased on the bridge^{3–6}, may be a contributing electron-transfer pathway in **3–5**. Thus in equation (1), effectively $k_2 \gg k_1$ for **3–5** leading to formation of $\text{D}^+\text{B}^-\text{A}$, which proceeds rapidly to D^+BA^- with k_3 . The measured CS rate constants monitor the slightly endothermic rate-determining step ${}^1\text{DBA} \rightarrow \text{D}^+\text{B}^-\text{A}$. There is no evidence from either the transient spectra or the CS kinetics for a significant build-up of $\text{D}^+\text{B}^-\text{A}$ population during the course of the reaction; that is, $k_3 \gg k_2$. This is reasonable in view of the fact that the reaction $\text{D}^+\text{B}^-\text{A} \rightarrow \text{D}^+\text{BA}^-$ is strongly exothermic, and therefore is expected to be very fast. The mono-exponential nature of the observed CS kinetics indicates also that $k_2 \gg k_4$.

The CR mechanism may contain contributions from both hole and electron transfer. The LUMO–LUMO energy gap between PI and the PPV bridges within **1–5** is ≥ 1 eV, so that CR via electron transfer from PI⁺ to D^+ should proceed slowly by the superexchange mechanism. In addition, the HOMO–HOMO energy gap between TET and the PPV bridges within **1–4** is ≥ 0.2 eV, so that CR via hole transfer should also be dominated by superexchange; that is $k_5 \gg k_6$ in equation (2).



On the other hand, the HOMO–HOMO energy gap between TET and the PPV bridge within **5** is sufficiently small to make hole injection from TET^+ into PPV5 kinetically competitive with superexchange ($k_5 \approx k_6$) by means of the additional CR mechanism $\text{D}^+\text{BA}^- \xrightarrow{k_6} \text{DB}^+\text{A}^- \xrightarrow{k_7} \text{DBA}$, where $k_7 \gg k_6$. The bi-exponential CR kinetics observed for **5** suggests that k_6 is comparable to k_8 . These results suggest that the lifetime of the D^+BA^- state could be significantly lengthened by increasing the HOMO–HOMO energy gap between D and B. This hole-injection CR mechanism is analogous to the electron-injection mechanism proposed for CS in **3–5**, and highlights the importance of the energy-matching requirements in promoting the molecular-wire behaviour of the bridge molecule.

We have produced a family of DBA molecules using oligomers of the conducting polymer PPV as the bridging motif. Photoinduced

charge separation in these systems is very weakly distance dependent, indicating that the unsaturated bridge acts as an incoherent molecular wire. If exponential distance dependence is assumed, the β value measured is amongst the lowest found to date. More importantly, we have presented experimental evidence for the critical role that the energy gap for donor-to-bridge charge injection plays in promoting molecular-wire behaviour in DBA molecules that possess highly conjugated bridges. Given that PPV polymers display unusually high electron and hole mobilities³¹, the ability to exercise photochemical control over charge injection into well-defined PPV oligomeric structures may lead to molecular electronic devices based on these materials. □

Received 22 June; accepted 14 August 1998.

- Barbara, P. F., Meyer, T. J. & Ratner, M. A. Contemporary issues in electron transfer research. *J. Phys. Chem.* **100**, 13148–13168 (1996).
- Wasielewski, M. R. Photoinduced electron transfer in supramolecular systems for artificial photosynthesis. *Chem. Rev.* **92**, 435–461 (1992).
- Ratner, M. A. & Jortner, J. in *Molecular Electronics* (eds Jortner, J. & Ratner, M.) 5–72 (Blackwell, Oxford, 1997).
- Davis, W. B., Wasielewski, M. R., Ratner, M. A., Mujica, V. & Nitzan, A. Electron transfer rates in bridged molecular systems: A phenomenological approach to relaxation. *J. Phys. Chem. A* **101**, 6158–6164 (1997).
- Okada, A., Chernyak, V. & Mukamel, S. Solvent reorganization in long-range electron transfer: density matrix approach. *J. Phys. Chem. A* **102**, 1241–1251 (1998).
- Pollard, W. T., Felt, A. K. & Friesner, R. A. The Redfield equation in condensed-phase quantum dynamics. *Adv. Chem. Phys.* **93**, 77–134 (1996).
- Winkler, J. R. & Gray, H. B. Electron transfer in ruthenium-modified proteins. *Chem. Rev.* **92**, 369–379 (1992).
- Moser, C. C., Keske, J. M., Warncke, K., Farid, R. S. & Dutton, P. L. Nature of biological electron transfer. *Nature* **355**, 796–802 (1992).
- Lewis, F. D. *et al.* Distance-dependent electron transfer in DNA hairpins. *Science* **277**, 673–676 (1997).
- Holmlin, R. E., Dandliker, P. J. & Barton, J. K. Charge transfer through the DNA base stack. *Angew. Chem. Int. Edn. Engl.* **36**, 2715–2730 (1998).
- Johnson, M. D., Miller, J. R., Green, N. S. & Closs, G. L. Distance dependence of intramolecular hole and electron transfer in organic radical ions. *J. Phys. Chem.* **93**, 1173–1176 (1989).
- Oevering, H. *et al.* Long-range photoinduced through-bond electron transfer and radiative recombination via rigid nonconjugated bridges: distance and solvent dependence. *J. Am. Chem. Soc.* **109**, 3258–3269 (1987).
- Helms, A., Heiler, D. & McLendon, G. Electron transfer in bis-porphyrin donor-acceptor compounds with polyphenylene spacers shows a weak distance dependence. *J. Am. Chem. Soc.* **114**, 6227–6238 (1992).
- Ribou, A.-C. *et al.* Intervalence electron transfer in pentaammineruthenium complexes of dipyrrolyl-porphyrins, dipyrrolylthiophene, and dipyrrolylfuran. *Inorg. Chem.* **33**, 1325–1329 (1994).
- Arrhenius, T. S., Blanchard-Desce, M., Dvornitzky, M., Lehn, J. M. & Malthete, J. Molecular devices: caroviologens as an approach to molecular wires—synthesis and incorporation into vesicle membranes. *Proc. Natl Acad. Sci. USA* **83**, 5355–5359 (1986).
- Wasielewski, M. R. *et al.* in *Photochemical Energy Conversion* (ed. Norris, J. R.) 135–147 (Elsevier, New York, 1989).
- Effenberger, F. & Wolf, H. C. Terminally substituted conjugated polyenes: synthesis and energy transfer properties. *New J. Chem.* **15**, 117–123 (1991).
- Tolbert, L. M. Solitons in a box: the organic chemistry of electrically conducting polyenes. *Acc. Chem. Res.* **25**, 561–568 (1992).
- Tour, J. M. Conjugated macromolecules of precise length and constitution. Organic synthesis for the construction of nanoarchitectures. *Chem. Rev.* **96**, 537–553 (1996).
- Grosshenny, V., Harriman, A. & Zissel, R. Towards the development of molecular wires: electron localization, exchange, and transfer in alkyne-bridged multinuclear complexes. *Angew. Chem. Int. Edn. Engl.* **34**, 2705–2708 (1996).
- Sachs, S. B. *et al.* Rates of interfacial electron transfer through π -conjugated spacers. *J. Am. Chem. Soc.* **119**, 10563–10564 (1997).
- Burgdorff, C., Ehrhardt, S. & Loehmannsroeben, H. G. Photophysical properties of tetracene derivatives in solution. 2. Halogenated tetracene derivatives. *J. Phys. Chem.* **95**, 4246–4249 (1991).
- Greenfield, S. R., Svec, W. A., Goszola, D. & Wasielewski, M. R. Multistep photochemical charge separation in rod-like molecules based on aromatic imides and diimides. *J. Am. Chem. Soc.* **118**, 6767–6777 (1996).
- Viehbeck, A., Goldberg, M. J. & Kovac, C. A. Electrochemical properties of polyimides and related imide compounds. *J. Electrochem. Soc.* **137**, 1460–1466 (1990).
- Perkampus, H.-H. *UV-VIS Atlas of Organic Compounds* (VCH, Weinheim, 1992).
- Marcus, R. A. On the theory of oxidation-reduction reactions involving electron transfer. I. *J. Chem. Phys.* **24**, 966–978 (1956).
- Mujica, V., Kemp, M., Roitberg, A. & Ratner, M. Current-voltage characteristics of molecular wires: Eigenvalue staircase, Coulomb blockade, and rectification. *J. Chem. Phys.* **104**, 7296–7305 (1996).
- Joachim, C. & Vinuesa, J. F. Length dependence of the electronic transparency (conductance) of a molecular wire. *Europhys. Lett.* **33**, 635–640 (1996).
- Samanta, M. P., Tian, W., Datta, S., Henderson, J. I. & Kubiak, C. P. Electronic conduction through organic molecules. *Phys. Rev. B: Condens. Matter* **53**, R7626–R7629 (1996).
- Burn, P. L. *et al.* Chemical tuning of the electronic properties of poly(p-phenylenevinylene)-based copolymers. *J. Am. Chem. Soc.* **115**, 10117–10124 (1993).
- Hoofman, R. J. O. M., De Haas, M. P., Siebbeles, L. D. A. & Warman, J. M. Highly mobile electrons and holes on isolated chains of the semiconducting polymer poly(phenylene vinylene). *Nature* **392**, 54–56 (1998).

Acknowledgements. This work was supported by the Division of Chemical Sciences, Office of Basic Energy Sciences, Department of Energy (M.R.W.), the Chemistry Division of the NSF and ONR (M.A.R.), and the Link Energy Foundation (W.B.D.).

Correspondence and requests for materials should be addressed to M.R.W. (e-mail: wasielew@chem.nyu.edu).

A large source of atmospheric nitrous oxide from subtropical North Pacific surface waters

John E. Dore^{*†}, Brian N. Popp^{†‡}, David M. Karl[†] & Francis J. Sansone[†]

^{*} Aquasearch, Inc., 73-4460 Queen Kaahumanu Highway, Suite 110, Kailua-Kona, Hawaii 96740, USA

[†] School of Ocean and Earth Science and Technology, Department of Oceanography, University of Hawaii, Honolulu, Hawaii 96822, USA

[‡] School of Ocean and Earth Science and Technology, Department of Geology and Geophysics, University of Hawaii, Honolulu, Hawaii 96822, USA

Nitrous oxide (N₂O), a trace gas whose concentration is increasing in the atmosphere, plays an important role in both radiative forcing and stratospheric ozone depletion^{1,2}. Its biogeochemical cycle has thus come under intense scrutiny in recent years. Despite these efforts, the global budget of N₂O remains unresolved, and the nature and magnitude of the sources and sinks continue to be debated^{3–5} despite the constraints that can be provided by characterizations of the gas^{6,7}. We report here the results of dual-isotope measurements of N₂O from the water column of the subtropical North Pacific Ocean. Nitrous oxide within the lower-euphotic and upper-aphotic zones is depleted in both ¹⁵N and ¹⁸O relative to its tropospheric and deep-ocean composition. These findings are consistent with a prediction, based on global mass-balance considerations, of a near-surface isotopically depleted oceanic N₂O source⁴. Our results indicate that this source, probably produced by bacterial nitrification, contributes significantly to the ocean-atmosphere flux of N₂O in the oligotrophic subtropical North Pacific Ocean. This source may act to buffer the isotopic composition of tropospheric N₂O, and is quantitatively significant in the global tropospheric N₂O budget. Because dissolved gases in near-surface waters are more readily exchanged with the atmospheric reservoir than those in deep waters, the existence of a quantitatively significant N₂O source at a relatively shallow depth has potentially important implications for the susceptibility of the source, and the ocean-atmosphere flux, to climatic influences.

Field work was carried out at the Hawaii Ocean Time-series Station ALOHA⁸ (22° 45' N, 158° W), a representative oligotrophic deep-ocean site. We collected seawater samples for isotopic analyses during the HOT-76 (September–October 1996), HOT-79 (January 1997) and HOT-82 (April 1997) cruises. Two archived seawater samples from the HOT-65 (September 1995) cruise were also analysed to ensure the efficacy of sample storage. We also analysed air samples collected during HOT-76 and from the roof of our Honolulu laboratory. Dissolved N₂O concentrations ranged from near air saturation (~6 nM) at the surface to a maximum of nearly 50 nM within the deep oxygen minimum at 700–800 m depth (Fig. 1). Concentration profiles were similar for HOT-76 and HOT-79, but showed some increased values below 100 m during HOT-82. Although the $\delta^{15}\text{N}$ and $\delta^{18}\text{O}$ values of dissolved N₂O display some variability between cruises, all three data sets show similar trends with depth: values near those of air at the surface, decreasing with depth to a minimum between 100 and 300 m, and then an increase with depth to 800 m (Fig. 1). The previously undescribed minima in the isotopic ratios differ from surface values by as much as 3.3‰ for $\delta^{15}\text{N}$ and 5.7‰ for $\delta^{18}\text{O}$. Because both isotopic ratios are lighter within this zone than above or below, the N₂O in this zone cannot result from a simple mixture of atmospheric N₂O with N₂O from deep water. The isotopic-minimum layer is located below the local salinity maximum and well

above the salinity minimum, so it seems not to be associated with either laterally advective hydrographical feature entering the region. Therefore, either production of light N_2O or preferential consumption of heavy N_2O occurs within this zone.

Consumption of N_2O at the isotopic-minimum zone is unlikely to be a significant process. Anoxic microsites can exist within particles and could be sites of active reduction of N_2O (ref. 9). However, a large kinetic isotope effect is associated with N_2O reduction, which should result in a residual pool of N_2O enriched in ^{15}N and ^{18}O (refs 10, 11). We observe an isotopic minimum, not a maximum. On the other hand, nitrification is an aerobic process, occurring throughout the ocean and characteristically exhibiting maximum rates within the lower-euphotic and upper-aphotic zones^{12,13}. Recent rate measurements at station ALOHA document active nitrification at the base of the euphotic zone¹³. The large negative fractionation factors associated with N_2O production via nitrification^{14,15} are consistent with the observation of N_2O isotopic minima within this layer. Although other metabolic pathways for N_2O production cannot be ruled out, the correlation of measurable nitrification activity with isotopically depleted N_2O strongly suggests nitrification as the responsible process.

We can estimate the magnitude and isotopic composition of this near-surface N_2O source as follows. During the late summer (HOT-76), low values of both $\delta^{15}N$ and $\delta^{18}O$ are found within the nitrifying layer, but with the development of more vigorous winter mixing (HOT-79 and HOT-82), the minima in the isotopic profiles are partially eroded, through exchange with the atmospheric N_2O above and perhaps also with the deep N_2O below (Fig. 1). In the absence of deep winter mixing, the isotopic composition of N_2O immediately below the euphotic zone will approximate that of a mixture of the isotopically enriched N_2O diffusing upwards across the thermocline and the isotopically depleted N_2O produced locally. The vertical flux of N_2O from deep water and its isotopic signature can be calculated using one-dimensional diffusion models^{16,17} and the measured gradients of N_2O concentration and N and O isotopes through the thermocline. These calculations yield an upward N_2O flux of $0.26 \pm 0.10 \mu\text{mol m}^{-2} \text{d}^{-1}$ with $\delta^{15}N = 12.0\text{‰}$ and $\delta^{18}O = 28.6\text{‰}$. The error in the magnitude of the flux results from the $\pm 40\%$ uncertainty in the vertical eddy-diffusion coefficient for this region¹⁶, while the calculation of the isotopic composition of the flux is independent of the value of this coefficient. At the isotopic minimum (215 m) during HOT-76, N_2O has a minimum composition of $\delta^{15}N = 5.8\text{‰}$ and $\delta^{18}O = 17.3\text{‰}$. Because we model this composition as a two-component mixture, and we now have

Table 1 Constraints on N_2O source strength at station ALOHA

Fraction of N_2O contributed to mixture by shallow source*	$\delta^{15}N$ of shallow N_2O source (‰ vs air- N_2)†	$\delta^{18}O$ of shallow N_2O source (‰ vs air- O_2)†	Magnitude of shallow N_2O source ($\mu\text{mol m}^{-2} \text{d}^{-1}$)‡	Magnitude of net sea-air N_2O flux ($\mu\text{mol m}^{-2} \text{d}^{-1}$)‡
95%	5.5	16.7	4.9 ± 2.0	5.2 ± 2.1
90%	5.1	16.0	2.3 ± 0.9	2.6 ± 1.0
80%	4.3	14.5	1.0 ± 0.4	1.3 ± 0.5
70%	3.1	12.5	0.6 ± 0.2	0.9 ± 0.3
60%	1.7	9.8	0.4 ± 0.2	0.7 ± 0.3
50%	-0.4	6.0	0.3 ± 0.1	0.5 ± 0.2
40%	-3.5	0.4	0.2 ± 0.1	0.4 ± 0.2
30%	-8.7	-9.1	0.1 ± 0.0	0.4 ± 0.1
20%	-19.0	-27.9	0.1 ± 0.0	0.3 ± 0.1
10%	-50.0	-84.4	0.0 ± 0.0	0.3 ± 0.1
5%	-112.0	-197.4	0.0 ± 0.0	0.3 ± 0.1

*We model the isotopic composition of N_2O at the isotopic minimum ($\delta^{15}N = 5.8\text{‰}$, $\delta^{18}O = 17.3\text{‰}$) as a two-component mixture of N_2O diffusing upwards from deep water ($\delta^{15}N = 12.0\text{‰}$, $\delta^{18}O = 28.6\text{‰}$) and N_2O produced locally by the shallow source.

†A given fractional contribution of the shallow N_2O source to the mixture yields these possible values for the isotopic composition of the shallow source.

‡The magnitudes of the shallow source and the net sea-air flux are calculated for a given fractional contribution using a vertical flux from deep water of $0.26 \pm 0.10 \mu\text{mol m}^{-2} \text{d}^{-1}$.

estimated the magnitude and composition of one component, knowledge of the relative proportions of the two sources will allow us to estimate the magnitude and isotopic composition of the second component.

The isotopic data can be used to constrain the fractional contribution of the shallow N_2O source to the N_2O pool at the isotopic minimum. One can see that if this fractional contribution is small, then the N_2O produced by the shallow source must be extremely depleted in the heavy isotopes (Table 1). The lowest $\delta^{18}O$ value ever reported for N_2O was -2.6‰ for a sample collected from a waste water treatment facility in which nitrification rates were extremely high¹⁵. It is very unlikely that N_2O produced at our isotopic minimum could be formed with a more depleted composition than the N_2O from this extreme environment. This lower limit for $\delta^{18}O$ implies that the shallow N_2O source must contribute $\geq 40\%$ of the N_2O to the pool at the isotopic minimum (Table 1). A similar argument can be made on the basis of extrema of $\delta^{15}N$ values. In the absence of other undescribed sources and/or sinks for N_2O in the water column at station ALOHA, the net flux from the ocean to the atmosphere will be the sum of the vertical flux from deep water and the production of N_2O at the isotopic minimum. The isotopic data suggest therefore that the net sea-air flux of N_2O at station ALOHA is at least $0.4 \pm 0.2 \mu\text{mol m}^{-2} \text{d}^{-1}$ (Table 1), which compares favourably with an estimate—based on global expedition data on

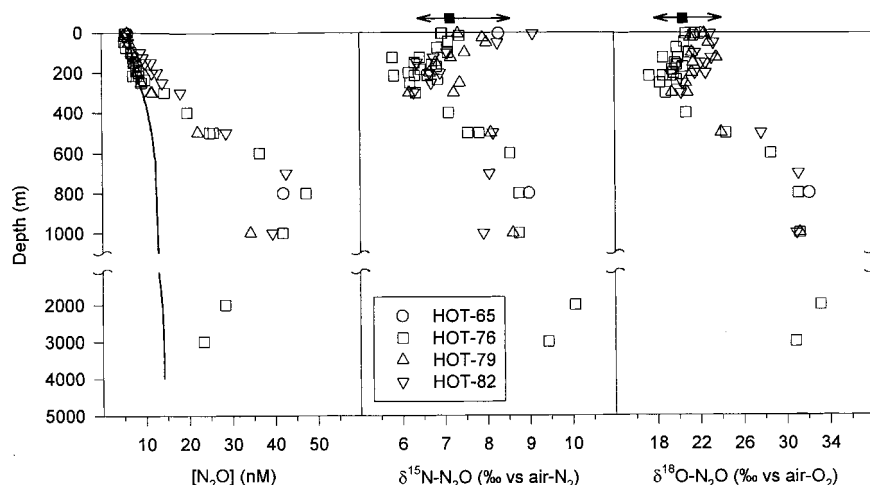


Figure 1 Isotope data from station ALOHA. Shown are depth profiles of concentration (left panel), nitrogen-isotope composition (centre panel) and oxygen-isotope composition (right panel) of N_2O at station ALOHA during four HOT cruises. Solid line in left panel indicates hypothetical N_2O concentration at

100% air saturation. Filled squares above the centre and right panels indicate measured isotopic composition of marine tropospheric N_2O during HOT-76. Arrows indicate ranges of measured isotopic compositions of tropospheric N_2O (ref. 6).

sea surface N_2O saturation state—of $0\text{--}0.8\ \mu\text{mol m}^{-2}\text{ d}^{-1}$ (ref. 18). If the shallow source actually contributes as much as 95% of the N_2O to the sea–air flux¹³, then the net flux to the atmosphere may be as high as $5.2 \pm 2.1\ \mu\text{mol m}^{-2}\text{ d}^{-1}$.

Kim and Craig⁴ suggested that the stable-isotope composition of tropospheric N_2O results from the mixing of three main endmembers: tropical soil N_2O emissions, the return flux of N_2O from the stratosphere, and an hypothesized yet undocumented near-surface oceanic N_2O source. Employing a global isotopic mass-balance model, they predicted that this latter source would have an approximate isotopic composition of 5‰ and 15‰ for $\delta^{15}\text{N}$ and $\delta^{18}\text{O}$, respectively. Although it has recently been suggested that the isotopic values these authors placed on the stratospheric end member may have been slight overestimates¹⁹, their predicted signature for the oceanic source would be little affected by a small downward revision of these figures. Our analyses provide direct evidence for their hypothesized oceanic N_2O source. Furthermore, the isotopic composition of our source compares very well with their prediction if this source contributes about 70–90% to the net ocean–atmosphere N_2O flux for this region (Table 1). Nevertheless, even with the isotopically depleted composition of the sea–air flux described here, the tropospheric N_2O budget cannot be balanced unless either the magnitude of this flux is much greater than estimated, or additional undescribed sources of isotopically depleted N_2O exist⁴.

A number of investigators have suggested that oceanic regions underlain by suboxic waters may represent an important source of N_2O to the atmosphere^{20–22}. Dual-isotope data have been obtained from two of these regions, the eastern tropical North Pacific and the

Arabian Sea^{23,24}. The results revealed that N_2O in the suboxic waters was highly enriched in both ^{15}N and ^{18}O , suggesting a denitrification-related source. However, the data from the suboxic zones fall well to the opposite side of the soil-gas/stratospheric mixing line from the tropospheric values (Fig. 2). These low-oxygen layers therefore cannot represent a significant source of N_2O to the troposphere, unless balanced by a yet-undescribed source. On the other hand, in both of these regions N_2O above the suboxic zone, within a layer of active nitrification near the base of the euphotic zone (75–150 m; refs 20, 22), as well as N_2O at the surface of an active upwelling zone in the Arabian Sea²⁴, are depleted in ^{15}N and only slightly enriched in ^{18}O with respect to tropospheric N_2O (refs 10, 23, 24; Fig. 2). Because this shallower N_2O more readily exchanges with the atmosphere than N_2O from the deeper suboxic sources, it is likely that the actual isotopic composition of the net N_2O flux to the atmosphere in these regions is not far from that of tropospheric N_2O . Hence, the central North Pacific, the Arabian Sea and the eastern tropical North Pacific may all serve to buffer the troposphere against the enriched isotopic composition of the stratospheric return flux and the depleted composition of the soil gas emissions.

The possibility that near-surface nitrification is one of the chief contributors to the global ocean–atmosphere flux of N_2O has several important implications. First, if sea–air N_2O fluxes are dependent not so much on slow diffusion across the thermocline but more upon production and recycling (that is, nitrification) of organic matter above the thermocline, then significant variability in N_2O flux can be expected at timescales on the order of days–weeks. Second, if the primary source of oceanic N_2O is near the surface rather than in deep waters, variability in wind-driven mixing, such as can be expected in response to climate change, will have a profound effect on the regional and global ocean–atmosphere fluxes of this important trace gas. Last, significant climatic phenomena such as El Niño may produce opposing effects on N_2O flux in different regions of the ocean. For example, while suppression of upwelling by El Niño may reduce N_2O in the surface waters of the equatorial Pacific²⁵, ecosystem shifts in the subtropical North Pacific in response to El Niño (which favour the dominance of nitrogen-fixing cyanobacteria) would be expected to increase nitrification rates in the upper water column, hence increasing near-surface N_2O production²⁶.

Although the data presented here provide direct evidence for an oceanic N_2O source consistent with the hypothesis of Kim and Craig⁴, we must be careful not to conclude that the tropospheric budget of N_2O is near closure. The temporal and spatial variability of N_2O fluxes and isotopic composition need to be better resolved before robust conclusions can be drawn. Few stable-isotope data on N_2O exist, and even fewer dual-isotope data, largely because of analytical limitations. Our analytical method allows for accurate and precise analysis of small-volume aquatic samples across nearly the entire concentration range found in nature, as well as air samples. This method also opens up the possibility of determining the microbiological pathways of N_2O production in natural aquatic systems by performing true tracer-level incubation experiments with stable isotopically-labelled compounds. Such studies should be able to elucidate the rates of N_2O production at specific sites, and also to yield information on the metabolic pathways involved. □

Methods

Isotopic analyses of N_2O . Twelve-litre Niskin bottles on a rosette system were used for seawater sample collection at up to 24 depths per cast. Subsamples for isotopic analyses were taken without aeration from the Niskin bottles into 250-ml serum vials, preserved with 1-ml saturated mercuric chloride solution, and capped with a butyl rubber stopper and aluminum seal for later analysis. Air samples were collected in 250-ml evacuated glass bulbs while facing into the prevailing wind. The system used for purging the gases from sea water and separating the gas of interest was originally designed for analysis of $\delta^{13}\text{C}$ of

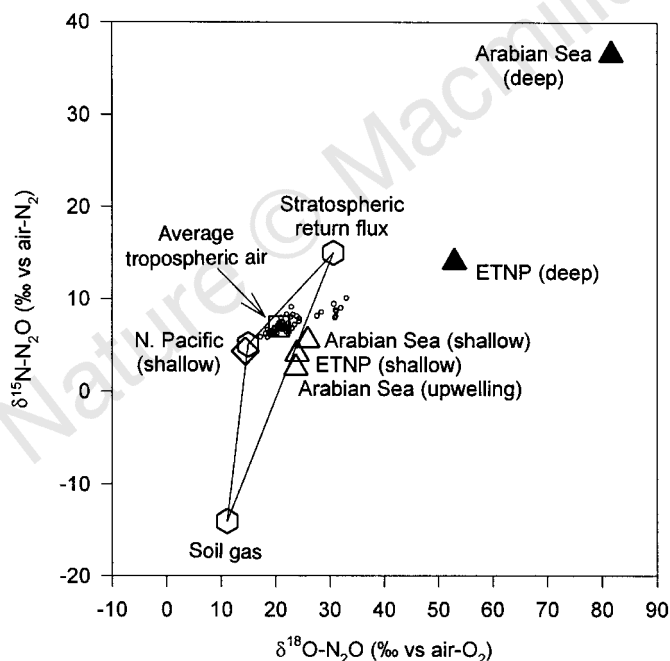


Figure 2 Property–property plot of oxygen- and nitrogen-isotope composition of N_2O . Small circles indicate water-column data from the subtropical North Pacific Ocean (this study). Hexagons indicate the three global endmembers for the tropospheric isotopic mass balance of N_2O : soil gas emissions, the stratospheric return flux, and the hypothesized near-surface oceanic source⁴. The diamond indicates the composition of the shallow North Pacific endmember revealed in this study, assuming that this source is responsible for 80% of the net sea–air N_2O flux at station ALOHA. The square indicates the average composition of tropospheric air⁶. Filled triangles show compositions of N_2O in deep suboxic layers from the eastern tropical North Pacific (ETNP) and the Arabian Sea²³. Open triangles show compositions of N_2O from shallow concentration maxima in the ETNP and the Arabian Sea²³, and from the surface waters of an active upwelling zone in the Arabian Sea²⁴.

dissolved methane in sea water²⁷. Briefly, sea water was transferred directly from the serum vial to a sparging column, and the dissolved gases stripped with ultra-pure He for 12 min. Air samples were flushed into the sparger by introduction into the carrier stream via the standard loop upstream of the sparger. The sparged gases were passed through columns containing Drierite and Ascarite to remove excess water vapour and CO₂, respectively, and the remaining condensable gases were trapped in a stainless-steel column immersed in liquid nitrogen. These gases were cryofocused²⁸, before separating N₂O from CO₂ and other trapped gases using a PorapLOT-Q (25 m × 0.32 mm) analytical column held at 10 °C. N₂O was then introduced directly into the ion source of a MAT 252 isotope-ratio-monitoring gas chromatograph/mass spectrometer (irmGC/MS). N₂O concentration and stable-isotope ratios (¹⁵R = ¹⁵N/¹⁴N and ¹⁸R = ¹⁸O/¹⁶O) were measured simultaneously by monitoring the ion currents of masses 44, 45 and 46 using the Finnigan ISODAT software. Nitrogen-isotope enrichments were corrected for the contribution of ¹⁷O to the mass 45 ion current^{29,30}. We report all isotopic values as per mil deviations with respect to atmospheric N₂ and O₂ using standard delta notation: δ(‰) = [(R_{sample}/R_{standard}) - 1] × 1,000. Our analytical scheme is capable of analysing as little as 1 nanomole of N₂O in less than 45 min. Analyses of samples in duplicate or triplicate yield reproducibilities of approximately ±0.3‰ for δ¹⁵N and ±0.7‰ for δ¹⁸O of dissolved N₂O in near-surface waters. We characterized the isotopic composition of our standard gas with respect to internationally accepted standards by the conversion of N₂O to N₂ plus CO₂ (refs 6, 31), followed by cryogenic separation and analysis of the gases produced using traditional methods. Calibration of N₂O concentration was achieved using a commercial gas mixture, and double-checked using dilutions of high-purity N₂O in He.

Received 31 December 1997; accepted 21 August 1998.

- Houghton, J. T., Jenkins, G. J. & Ephraums, J. J. (eds) *Climate Change: The IPCC Scientific Assessment* (Cambridge Univ. Press, 1990).
- Crutzen, P. J. The influence of nitrogen oxides on the atmospheric ozone content. *Q. J. R. Meteorol. Soc.* **96**, 320–325 (1970).
- Bouwman, A. F., Van der Hoek, K. W. & Olivier, J. G. J. Uncertainties in the global source distribution of nitrous oxide. *J. Geophys. Res.* **100**, 2785–2800 (1995).
- Kim, K.-R. & Craig, H. Nitrogen-15 and oxygen-18 characteristics of nitrous oxide: a global perspective. *Science* **262**, 1855–1857 (1993).
- McElroy, M. B. & Jones, D. B. A. Evidence for an additional source of atmospheric N₂O. *Glob. Biogeochem. Cycles* **10**, 651–659 (1996).
- Kim, K.-R. & Craig, H. Two-isotope characterization of N₂O in the Pacific Ocean and constraints on its origin in deep water. *Nature* **347**, 58–61 (1990).
- Zafriou, O. Laughing gas from leaky pipes. *Nature* **347**, 15–16 (1990).
- Karl, D. M. & Lukas, R. The Hawaii Ocean Time-series (HOT) program: background, rationale and field implementation. *Deep-Sea Res. II* **43**, 129–156 (1996).
- Allredge, A. L. & Cohen, Y. Can microscale chemical patches persist in the sea? Microelectrode study of marine snow, fecal pellets. *Science* **235**, 689–691 (1987).
- Yoshida, N., Hattori, A., Saino, T., Matsuo, S. & Wada, E. ¹⁵N/¹⁴N ratio of dissolved N₂O in the eastern tropical Pacific Ocean. *Nature* **307**, 442–444 (1984).
- Yamazaki, T., Yoshida, N., Wada, E. & Matsuo, S. N₂O reduction by *Azotobacter vinelandii* with emphasis on kinetic isotope effects. *Plant Cell Physiol.* **28**, 263–271 (1989).
- Ward, B. B. in *Nitrification* (ed. Prosser, J. I.) 157–184 (IRL, Oxford, 1986).
- Dore, J. E. & Karl, D. M. Nitrification in the euphotic zone as a source for nitrite, nitrate and nitrous oxide at Station ALOHA. *Limnol. Oceanogr.* **41**, 1619–1628 (1996).
- Yoshida, N. ¹⁵N-depleted N₂O as a product of nitrification. *Nature* **355**, 528–529 (1988).
- Yoshinari, T. & Wahlen, M. Oxygen isotope ratios in N₂O from nitrification at a wastewater treatment facility. *Nature* **317**, 349–350 (1985).
- Li, Y.-H., Peng, T.-H., Broecker, W. S. & Östlund, H. G. The average vertical mixing coefficient for the oceanic thermocline. *Tellus B* **36**, 212–217 (1984).
- Jørgensen, B. B. A theoretical model of the stable sulfur isotope distribution in marine sediments. *Geochim. Cosmochim. Acta* **43**, 363–374 (1979).
- Nevison, C. D., Weiss, R. F. & Erickson, D. J. Global oceanic emissions of nitrous oxide. *J. Geophys. Res.* **100**, 15809–15820 (1995).
- Rahn, T. & Wahlen, M. Stable isotope enrichment in stratospheric nitrous oxide. *Science* **278**, 1776–1778 (1997).
- Codispoti, L. A. & Christiansen, J. P. Nitrification, denitrification and nitrous oxide cycling in the eastern tropical South Pacific Ocean. *Mar. Chem.* **16**, 277–300 (1985).
- Law, C. S. & Owens, N. J. P. Significant flux of atmospheric nitrous oxide from the northwest Indian Ocean. *Nature* **346**, 826–828 (1990).
- Naqvi, S. W. A. & Noronha, R. J. Nitrous oxide in the Arabian Sea. *Deep-Sea Res.* **38**, 871–890 (1991).
- Yoshinari, T. et al. Nitrogen and oxygen isotopic composition of N₂O from suboxic waters of the eastern tropical North Pacific and the Arabian Sea—measurement by continuous-flow isotope-ratio monitoring. *Mar. Chem.* **56**, 253–264 (1997).
- Naqvi, S. W. A. et al. Budgetary and biogeochemical implications of N₂O isotope signatures in the Arabian Sea. *Nature* **394**, 462–464 (1998).
- Cline, J. D., Wisegarver, D. P. & Kelly-Hansen, K. Nitrous oxide and vertical mixing in the equatorial Pacific during the 1982–1983 El Niño. *Deep-Sea Res.* **34**, 857–873 (1987).
- Karl, D. M. et al. Ecosystem changes in the North Pacific subtropical gyre attributed to the 1991–92 El Niño. *Nature* **373**, 230–234 (1995).
- Sansone, F. J., Popp, B. N. & Rust, T. M. Stable carbon isotopic analysis of low-level methane in water and gas. *Anal. Chem.* **69**, 40–44 (1997).
- Popp, B. N., Sansone, F. J., Rust, T. M. & Merritt, D. A. Determination of concentration and carbon isotopic composition of dissolved methane in sediments and nearshore waters. *Anal. Chem.* **67**, 405–411 (1995).

- Craig, H. Isotopic standards for carbon and oxygen and correction factors for mass-spectrometric analysis of carbon dioxide. *Geochim. Cosmochim. Acta* **12**, 133–149 (1957).
- Brand, W. A. PRECON: a fully automated interface for the pre-GC concentration of trace gases in air for isotopic analyses. *Isotopes Environ. Health Stud.* **31**, 277–284 (1995).
- Yoshida, N. & Matsuo, S. Nitrogen isotope ratio of atmospheric N₂O as a key to the global cycle of N₂O. *Geochim. J.* **17**, 231–239 (1983).

Acknowledgements. We thank the members of the Hawaii Ocean Time-series (HOT) program and M. E. Holmes for sample collection, and T. Rust for technical support. This research was supported by the US National Science Foundation.

Correspondence and requests for materials should be addressed to J.D. (e-mail: asjdore@ilhawaii.net). Supporting data sets from station ALOHA are available at <http://hahana.soest.hawaii.edu>.

Dating topography of the Sierra Nevada, California, using apatite (U–Th)/He ages

Martha A. House*†, Brian P. Wernicke* & Kenneth A. Farley*

* Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA

The upward motion of rock masses relative to the Earth's surface has been documented for most of the main mountain belts using thermochronological and petrological techniques. More fundamental to the physical processes of mountain building, however, is the motion of the Earth's surface itself, which remains elusive¹. Here we describe a technique for estimating the age of topographic relief by mapping the low-temperature thermal structure imparted by river incision using the ages of apatites determined from their uranium, thorium and helium contents. The technique exploits horizontal variations in temperature in the shallow crust that result from range-normal river drainages^{2,3}, because cooling beneath ancient river valleys occurs earlier than beneath intervening ridges. Our results from the Sierra Nevada, California, indicate that two of the modern transverse drainages, the Kings and the San Joaquin, had developed deep canyons by the Late Cretaceous period, suggesting that the high topography of the range is ~50–60 million years older than generally thought^{4–6}.

Neglecting lateral variations in reduced heat flux, radioactive heat production, mean surface temperature and advection of heat due to erosion, the temperature *T* in the shallow crust below a periodic topography is given by:

$$T = T_0 + \frac{q_m z}{k} + \frac{\rho H_s h_r}{k} (1 - e^{-z/h_r}) + \left(\beta - \frac{q_m}{k} - \frac{\rho H_s h_r}{k} \right) h_0 \cos \frac{2\pi x}{\lambda} e^{-2\pi z/\lambda} \quad (1)$$

where *x* and *z* are horizontal distance and depth below mean elevation, respectively, *h*₀ and *λ* are amplitude and wavelength of relief, respectively⁷, *ρH*_s and *q*_m are surface radioactive heat production and reduced heat flow, respectively, and the remaining variables are described in Fig. 1 legend. For a mountain belt with a geothermal gradient near 20 °C km⁻¹, lapse rate of surface temperature *β* of 4.5 °C km⁻¹ (ref. 8), and relief 2*h*₀ of 2 km, the maximum horizontal variation in temperature near *z* = *h* is ~30 °C. Because the thermal amplitude attenuates with characteristic depth *λ*/2*π*, cooling ages would be most sensitive to long-wavelength topography in the upper few kilometres of the crust.

The thermochronometer with the greatest potential for measuring this effect is the (U–Th)/He method on apatite⁹, which has a closure temperature near 75 °C (ref. 10). For this system, we wish to know what horizontal age variation, *Δt*, results from a given *h*₀ and

† Present address: Department of Earth & Atmospheric Sciences, Saint Louis University, St. Louis, Missouri 63103, USA.

λ , the two parameters that define the topography. For a thermal and erosional history relevant to the Sierra Nevada (Fig. 1), we obtain $\Delta t \sim 20$ Myr, using equation (1) and a numerical solution to the He production-diffusion equation¹¹ (using experimentally derived diffusion parameters for He in apatite¹⁰). Because uncertainties in $(U-Th)/He$ ages are typically $<5\%$ (or <5 Myr for samples with Cretaceous/Tertiary ages), age variations due to topography should be detectable. Age variation is strongly dependent on h_0 (Δt ranging from 3 to 40 Myr for canyons 0.4–6.0 km deep), and less sensitive to λ (Δt ranging from 20 to 27 Myr for wavelength of 20–70 km). Therefore, horizontal age profiles should constrain the palaeotopography, wherein observed periodicity in age would yield λ , the oldest ages would yield a minimum age for river incision, and Δt would be used to estimate h_0 , provided that the geotherm and cooling history could be estimated.

Several important effects would damp thermal topography, leading to an underestimate of relief $2h_0$ for a given age contrast

Δt . One is advection, which is not likely to be important in the slowly eroding Sierra Nevada, but could be important in other mountain belts. A second is cooling from the eastern or western edges of the Sierra Nevada proper, where incision would be near zero and isotherms more planar. A third includes groundwater flow, wherein a regional aquifer transports heat from high areas to low areas¹².

To test the technique, we collected 36 samples for $(U-Th)/He$ dating along a 200-km range-parallel transect through the southern Sierra Nevada batholith, crossing the points of maximum valley incision at elevations near 2 km (Fig. 2). We chose the Sierra Nevada because it is a relatively simple west-tilted block. Its western margin has been near sea level since at least 80 Myr ago^{4,5} and there is no evidence of significant folding or faulting along the transect, or within the Great Valley sedimentary prism to the west. The heat production of Sierran plutons, which if strongly variable along the transect would compromise the technique, varies mainly in a direction normal to the range axis¹³. To obtain a nearly horizontal profile, samples near the modern valleys were collected from local summits and those from intervening ridges were collected from the bottoms of secondary valleys. To avoid any effects from local, small-volume Late Cenozoic igneous centres, we did not sample from within a few kilometres of them. Samples were taken from below the upper 10 cm of the bedrock surface to avoid transient thermal effects including lightning, forest fires and afternoon heating¹¹.

Total erosion of this part of the range since the end of granitic plutonism at 80–85 Myr ago is ~ 6 –8 km (ref. 14), yielding an average erosion rate of ~ 0.10 mm yr^{-1} . This rate is broadly consistent with the incision rate of the San Joaquin River since the deposition (~ 10 Myr ago) of fluvial gravels along the western flank of the range¹⁰, which extend eastwards to within 10 km of our sample traverse⁴, and with unroofing rates estimated from the tilt history of Cenozoic sediments in the Great Valley⁷. Age–elevation profiles from the Kings and Merced drainages based on apatite and sphene fission-track data indicate rapid cooling from 270 to 100 °C from 75 to 65 Myr ago¹⁵. In contrast, $(U-Th)/He$ ages from the same apatite separates yielded ages ranging from 30 to 75 Myr with a well defined elevation gradient of ~ 20 Myr km^{-1} (ref. 16). These data, combined with modern heat-flow measurements, indicate low to moderate geothermal gradients (10 – 25 °C km^{-1}) in the uppermost crust since ~ 70 Myr ago^{13,15,16}.

When projected onto the line of section in Fig. 2, the age profile is periodic with $\lambda \approx 70$ km and $\Delta t \approx 20$ –30 Myr, crudely mirroring

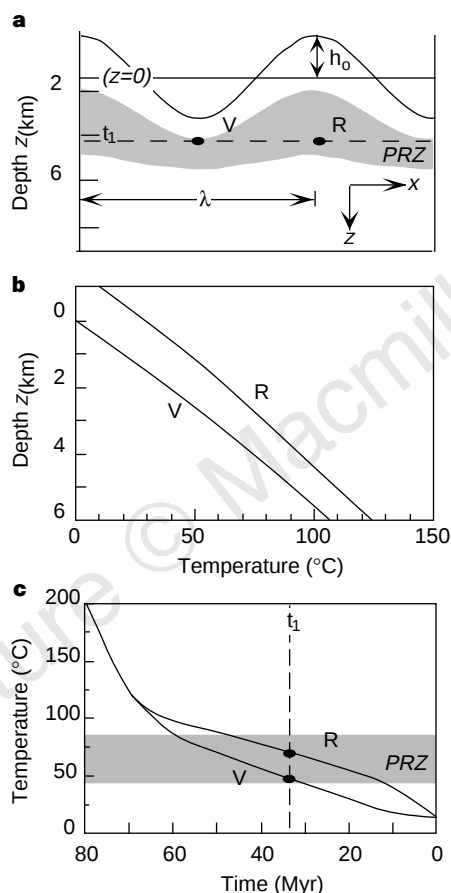


Figure 1 Thermal history of rock samples below periodic topography. **a**, Schematic range-parallel cross-section showing the influence of topography on isotherms bounding the partial retention zone (PRZ, shaded) for helium in apatite (45–85 °C; ref. 11). At time t_1 , the sample below the valley (V) is nearly closed to helium diffusion, while the sample below the ridge (R) remains open, resulting in a younger $(U-Th)/He$ age beneath the ridge. **b**, Steady-state geotherms beneath valley and ridge sites described by equation (1) using nominal central Sierran heat-flow parameters¹³ of reduced heat flow q_m (21 mW m^{-1}), characteristic depth of heat production h_r (10 km), thermal conductivity k (2.4 W $K^{-1} m^{-1}$), surface radioactive heat production ρH_s (1 $\mu W m^{-3}$), temperature at depth $z = 0$ km T_0 (15 °C), and lapse rate of mean surface temperature β (4.5 °C km^{-1}). **c**, Hypothetical cooling histories of valley and ridge samples. Rapid cooling trajectory above ~ 100 °C based on higher temperature thermochronometers for central Sierra^{15,16}. Trajectory through PRZ assumes unroofing at a constant rate of 0.08 mm yr^{-1} of a steady-state topography with $h_0 = 1$ km and $\lambda = 70$ km from a depth $z = 6$ km at 80 Myr, using geotherms from **b**.

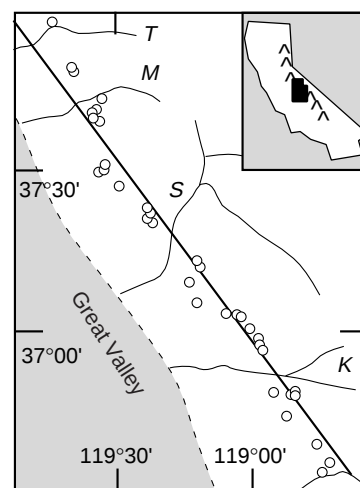


Figure 2 Study area. Shown are sample locations (open circles) and position of topographic profile (heavy line) in the Sierra Nevada, spanning the Tuolumne (T), Merced (M), San Joaquin (S) and Kings (K) drainages. The boundary between San Joaquin Valley (grey) and Sierra Nevada is indicated by the dashed line. Inset, location of Sierran crest (inverted 'vee' symbols) in the state of California.

the modern topography (Fig. 3). The widths of the age peaks appear to correlate with the widths of the valleys, with a wider peak for the San Joaquin drainage than for the Kings, and no apparent age effect for the much smaller Merced and Tuolumne drainages. The age profile, if indeed the result of palaeotopography, suggests that the large canyons of the San Joaquin and Kings rivers were incised by 70–80 Myr ago, whereas the smaller Tuolumne and Merced drainages were either too small to have a thermal effect or had not yet been formed. Other explanations for the variation, including heat production contrasts, Late Cenozoic magmatism, or kilometre-scale folding or faulting normal to the range axis, are inconsistent with the structural and thermal history of the range, as discussed above.

According to our calculations relating age variations to topographic amplitude in the Sierras, the observed 20–30 Myr variation in helium ages correspond to long-wavelength relief of 2–4 km. A more precise estimate will be possible using more detailed numerical calculations and sensitivity analysis of the function $h_0(\Delta t)$ to various assumptions regarding the geotherm and unroofing history (M.A.H. *et al.*, manuscript in preparation). Whatever the precise estimate, it represents a minimum for the Late Cretaceous mean elevation at the crest of the range. First, by analogy with the modern high ranges, we expect the amplitude of short-wavelength topography ($\lambda < 20$ km) to have been at least 250 m (Fig. 3b), adding about 500 m to the total relief. Second, the crests of mountain ranges are substantially higher than ridge elevations where relief is maximum. For example, the crest of the modern Sierra lies 45 km northeast of our profile, with both mean and summit elevations ~30% higher than those along the profile. If we incorporate these factors into a Cretaceous elevation model with long-wavelength relief of 3 km (corresponding to Δt of 25 Myr), we infer mean elevation of the Sierran crest of ~4.5 km, a value similar to that of

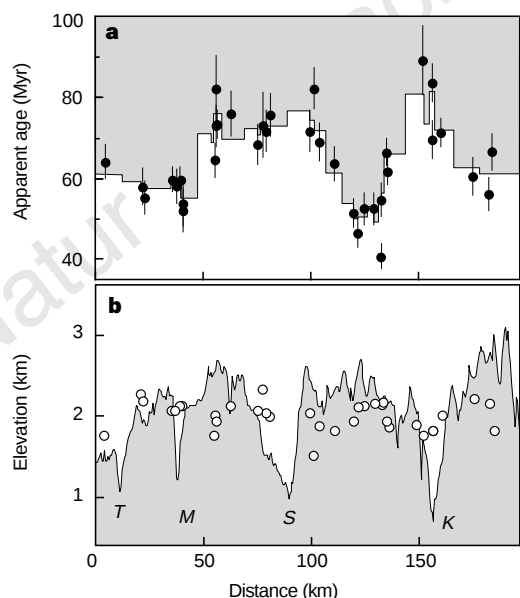


Figure 3 (U-Th)/He ages along range-parallel profile. **a**, Elevation-adjusted helium ages for samples projected onto profile in Fig. 2. Helium and U-Th analyses were performed on the same aliquots at Caltech following procedures in refs 16, 26. Samples of euhedral, inclusion-free igneous apatite from undeformed granitic plutons were analysed in replicate, with mean ages ranging from 44.5 to 84.6 Myr. Errors shown are 1 σ and reflect analytical uncertainties in helium and U-Th measurements^{16,26}. Ages were corrected for small differences in sample elevation above or below 2,000 m using the observed age-elevation gradient¹⁶. Stepped curve is defined by three-sample average of elevation-adjusted ages. A complete table of analytical results is available; see Supplementary Information. **b**, Location and elevation of samples projected onto profile in Fig. 2, plotted with topography along profile. Drainages are labelled as in Fig. 2.

the modern central Andes.

Cretaceous relief may be estimated independently of Δt by considering the difference between erosion of the valleys after initial incision and the total erosion since the Cretaceous. To preserve their old ages, the valley samples must have cooled below 50 °C by 70–80 Myr ago, implying 2–4 km of valley erosion above the 2,000-m profile since that time (assuming a geothermal gradient of 10–25 °C km⁻¹, ref. 16, Fig. 3b). Provided that total erosion of the central Sierra since 80 Myr ago was 6–8 km, and the ridges were eroding at a nominal rate of ~0.1 mm yr⁻¹ during the late Cretaceous (1 km of erosion between 70 and 80 Myr ago), then the ridges would have stood ~3 km above the valleys at ~70 Myr ago, similar to the estimate based on the age variations.

The modern relief is ~1 km along the profile (Fig. 3b) and the mean elevation of the crest of the range is¹⁷ 2.8 km, implying that (1) the average erosion rate of the ridges was roughly twice that of the valleys since ~70 Myr ago, leading to an overall reduction in Cenozoic relief, and (2) that mean elevation has decreased since 70 Myr ago. A model of decreasing mean elevation and relief since the Cretaceous challenges the long-standing hypothesis, based on palaeobotanical and geomorphological arguments, that the Sierra were a low-standing range of subdued relief for most of early Tertiary time, followed by rapid uplift to its current mean crestal elevation in the past 10–20 Myr (refs 4, 6). However, reinterpretation of the geomorphic data in terms of a flexural-isostatic uplift model is consistent with progressive lowering of mean elevation since the Cretaceous¹⁸, and the profound buoyancy loss of the crust since 20 Myr ago due to crustal thinning poses difficulties¹⁹. Further, the basis of the earlier palaeobotanical interpretations is in dispute²⁰, and recent work on floras in western Nevada, previously thought to indicate uplift over the past 10 million years, suggest a loss of some 1.5 km in mean elevation between 16 Myr ago and the present²¹.

If our inferences regarding Late Cretaceous relief are to be reconciled with a low-elevation, low-relief Sierra in the early Tertiary, it would require reduction of mean elevation of ~2 km in just 10–20 million years. Such a reduction could not be purely erosional, because isostatic considerations require at least 10–15 km of erosion, more than double the total post-85-Myr erosion. The only likely tectonic explanation for a significant post-batholithic, pre-extensional buoyancy loss would be the generation of the thick column of upper-mantle eclogitic rocks known from xenolith studies²² at the base of the cooling Sierran batholith. However, the age of crystallization of the eclogitic rocks is 80–120 Myr (ref. 23), too early to account for significant post-batholithic lowering²³.

We conclude that the Sierra developed an Andean-scale topography between ~185 Myr (the youngest marine strata preserved in batholith wallrocks²⁴) and 70–80 Myr (the minimum age of deep incision), coinciding with crustal thickening via arc magmatism in the Sierra and contractional deformation to the east. By 70–80 Myr ago, both the San Joaquin and Kings drainages had deeply incised the western flank of the range, after which we infer gradual reduction of mean crestal elevation from ~4.5 km to 2.8 km and of long-wavelength relief from ~3 km to 1 km. Although the buoyancy loss from crustal thinning since 20 Myr would presumably have resulted in accelerated loss of mean elevation, the synchronous loss of a cold, eclogitic root and its replacement with hot peridotite, especially near the crest of the range^{19,22,25}, may have added comparable or greater buoyancy. So it remains an open question whether these tectonic influences on buoyancy augmented or diminished the loss of mean elevation due to erosion since 20 Myr ago (ref. 18). □

Received 10 December 1997; accepted 27 July 1998.

- England, P. C. & Molnar, P. Surface uplift, uplift of rocks and exhumation of rocks. *Geology* **18**, 1173–1177 (1990).
- Mancktelow, N. S. & Grasemann, B. Time-dependent effects of heat advection and topography on cooling histories during erosion. *Tectonophysics* **270**, 167–195 (1997).

3. Stüwe, K., White, L. & Brown, R. The influence of eroding topography on steady-state geotherms. Application to fission track analysis. *Earth Planet. Sci. Lett.* **124**, 63–74 (1994).
4. Huber, N. K. Amount and timing of late Cenozoic uplift and tilt of the central Sierra Nevada, California—Evidence from the upper San Joaquin River basin. *US Geol. Surv. Prof. Pap.* **1197**, 1–28 (1981).
5. Unruh, J. R. The uplift of the Sierra Nevada and implications for late Cenozoic epeirogeny in the western Cordillera. *Geol. Soc. Am. Bull.* **103**, 1395–1404 (1991).
6. Axelrod, D. I. In *Integrated Earth and Environmental Evolution of the Southwestern U.S.* (eds Ernst, W. G. & Nelson, C. A.) 70–79 (Bellweather, Columbia, MD, 1998).
7. Turcotte, D. L. & Schubert, G. *Geodynamics* (Wiley, New York, 1982).
8. Birch, F. Flow of heat in the Front Range, Colorado. *Geol. Soc. Am. Bull.* **61**, 567–630 (1950).
9. Zeitler, P. K., Herczeg, A. L., McDougall, I. & Honda, M. U-Th-He dating of apatite: a potential thermochronometer. *Geochim. Cosmochim. Acta* **51**, 2865–2868 (1987).
10. Wolf, R. A., Farley, K. A. & Silver, L. T. Helium diffusion and low temperature thermochronometry of apatite. *Geochim. Cosmochim. Acta* **60**, 4231–4240 (1996).
11. Wolf, R. A., Farley, K. A. & Kass, D. M. Modeling of the temperature sensitivity of the apatite (U-Th)/He thermochronometer. *Chem. Geol.* **148**, 105–114 (1998).
12. Lachenbruch, A. H. & Sass, J. H. In *The Earth's Crust: Its Nature and Physical Properties* (eds Heacock, J. G. et al.) 626–675 (Am. Geophys. Union, Washington DC, 1977).
13. Saltus, R. W. & Lachenbruch, A. H. Thermal evolution of the Sierra Nevada: implications of new heat flow data. *Tectonics* **10**, 325–344 (1991).
14. Ague, J. J. & Brimhall, G. H. Magmatic arc asymmetry and distribution of anomalous plutonic belts in the batholiths of California: Effects of assimilation, crustal thickness, and depth of crystallization. *Geol. Soc. Am. Bull.* **100**, 912–927 (1988).
15. Dumitru, T. A. Subnormal Cenozoic geothermal gradients in the extinct Sierra Nevada magmatic arc: consequences of Laramide and post-Laramide shallow-angle subduction. *J. Geophys. Res.* **95**, 4925–4941 (1990).
16. House, M. A., Wernicke, B. P., Farley, K. A. & Dumitru, T. A. Cenozoic thermal evolution of the central Sierra Nevada, CA from (U-Th)/He thermochronometry. *Earth Planet. Sci. Lett.* **151**, 167–179 (1997).
17. Diment, W. H. & Urban, T. C. Average elevation map of the conterminous United States (Gilluly averaging method). *US Geol. Surv. Geophys. Invest. Map* **933**, (1981).
18. Small, E. E. & Anderson, R. S. Geomorphically driven Late Cenozoic rock uplift in the Sierra Nevada, California. *Science* **270**, 277–280 (1995).
19. Wernicke, B. et al. Origin of high mountains in the continents: the southern Sierra Nevada. *Science* **271**, 190–193 (1996).
20. Wolfe, J. A. An analysis of present-day terrestrial lapse rates in western conterminous United States and their significance to paleoaltitudinal estimates. *US Geol. Surv. Bull.* **1964**, 35 (1992).
21. Wolfe, J. A., Schorn, H. E., Forest, C. E. & Molnar, P. Paleobotanical evidence for high altitudes in Nevada during the Miocene. *Science* **276**, 1672–1675 (1997).
22. Ducea, M. N. & Saleeby, J. B. Buoyancy sources for a large, unrooted mountain range, the Sierra Nevada, California: Evidence from xenolith thermobarometry. *J. Geophys. Res.* **101**, 8229–8244 (1996).
23. Ducea, M. N. & Saleeby, J. B. The age and origin of a thick mafic-ultramafic root from beneath the Sierra Nevada batholith: Part I: Geochronology. *Contrib. Mineral. Petrol.* (in the press).
24. Saleeby, J. B. & Busby-Spera, C. In *The Cordilleran Orogen: Conterminous U.S.* (eds Burchfiel, B. C., Lipman, P. W. & Zoback, M. L.) 107–168 (Geol. Soc. Am. DNAG G3, Boulder, 1992).
25. Park, S. K., Hirasuna, B., Jiracek, G. R. & Kinn, C. Magnetotelluric evidence of lithospheric mantle thinning beneath the southern Sierra Nevada. *J. Geophys. Res.* **101**, 16241–16255 (1996).
26. Farley, K. A., Wolf, R. A. & Silver, L. T. The effects of long alpha-stopping distances on (U-Th)/He dates. *Geochim. Cosmochim. Acta* **60**, 4223–4230 (1996).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements. This work is part of the Southern Sierra Nevada Continental Dynamics Project, supported by NSF's Continental Dynamics Program. We thank M. Ducea and J. Saleeby for permission to discuss their results before publication.

Correspondence and requests for materials should be addressed to M.A.H. (e-mail: house@eas.slu.edu).

Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity

Marcel G. A. van der Heijden^{†‡}, John N. Klironomos^{†‡}, Margot Ursic[‡], Peter Moutoglou[§], Ruth Streitwolf-Engel^{*}, Thomas Boller^{*}, Andres Wiemken^{*} & Ian R. Sanders^{*}

^{*} Botanisches Institut der Universität Basel, Hebelstrasse 1, 4056 Basel, Switzerland

[‡] Department of Botany, University of Guelph, Ontario N1G 2W1, Canada

[§] Premier Tech Rivière-du-Loup, Québec G5R 4C9, Canada

[†] These authors contributed equally to this work

The functioning and stability of terrestrial ecosystems are determined by plant biodiversity and species composition^{1–5}. However, the ecological mechanisms by which plant biodiversity and species composition are regulated and maintained are not well understood. These mechanisms need to be identified to ensure successful management for conservation and restoration of

diverse natural ecosystems. Here we show, by using two independent, but complementary, ecological experiments, that below-ground diversity of arbuscular mycorrhizal fungi (AMF) is a major factor contributing to the maintenance of plant biodiversity and to ecosystem functioning. At low AMF diversity, the plant species composition and overall structure of microcosms that simulate European calcareous grassland fluctuate greatly when the AMF taxa that are present are changed. Plant biodiversity, nutrient capture and productivity in macrocosms that simulate North American old-fields increase significantly with increasing AMF-species richness. These results emphasize the need to protect AMF and to consider these fungi in future management practices in order to maintain diverse ecosystems. Our results also show that microbial interactions can drive ecosystem functions such as plant biodiversity, productivity and variability.

The mechanisms that control plant biodiversity are still being debated. The ability of many plant species to co-exist, and thus to determine plant biodiversity, can be explained by competitive interactions^{6,7}, by spatial or temporal resource partitioning^{8,9}, by disturbance creating new patches for colonization^{10,11}, and by interactions among different functional groups of organisms that constitute ecosystems^{12–15}. So far, little attention has been paid to the effects of microbe–plant interactions, particularly the mycorrhizal symbiosis, on ecosystem variability, productivity and plant biodiversity. AMF are abundant in soils of most ecosystems; these fungi form mutualistic symbiotic associations with the roots of ~80% of all terrestrial plant species, thereby acting as extensions of plant root systems and increasing nutrient uptake, especially of phosphorus¹⁶. The presence of AMF in ecosystems increases plant biodiversity¹⁷. However, not only are AMF present in most ecosystems, but communities of AMF also occur, which vary in species composition, species number and, therefore, in AMF biodiversity^{18,19}. Furthermore, AMF biodiversity is greatly reduced in some ecosystems²⁰. Different AMF species induce differential growth responses, in terms of biomass production and clonal growth patterns, of co-existing plant species of calcareous grasslands^{21,22}. On the basis of these results, we proposed that the species composition and diversity of AMF communities have the potential to determine plant biodiversity in natural ecosystems²².

Here we show for the first time, to our knowledge, that species composition and species richness of AMF is an important contributor to plant species composition, variability, productivity and biodiversity in artificial microcosms and macrocosms.

In a greenhouse experiment (experiment 1), we compared the effects of four different native AMF taxa, which were all isolated from the soil of a calcareous grassland, and of a combination of these four AMF taxa on the species composition and structure of 48 microcosms simulating European calcareous grassland. When we compared any of these AMF treatments with the non-mycorrhizal control treatment, we found that eight of the eleven plant species were almost completely dependent on the presence of AMF to be successful in the microcosms (Fig. 1; plant species: *Brachypodium pinnatum*, *Centaurium erythraea*, *Hieracium pilosella*, *Lotus corniculatus*, *Prunella grandiflora*, *Prunella vulgaris*, *Sanguisorba officinalis* and *Trifolium pratense*). In contrast, *Carex flacca*, the only plant species of the microcosms that does not have a symbiotic relationship with AMF, had the highest biomass in the non-mycorrhizal treatment (Fig. 1f). Furthermore, the biomass produced by most of the plant species varied significantly among treatments with different single AMF taxa (Fig. 1), indicating that different plant species benefited to different extents from different AMF taxa. For example, the biomass of *Sanguisorba officinalis* and of *Trifolium pratense* was highest in microcosms inoculated with AMF A; the biomass of *Brachypodium pinnatum* was highest in microcosms containing AMF C; the biomass of *Prunella vulgaris* was highest in microcosms containing AMF D (Fig. 1e, h–j). Although altering the AMF taxa in the soil had no significant effect on the biomass of

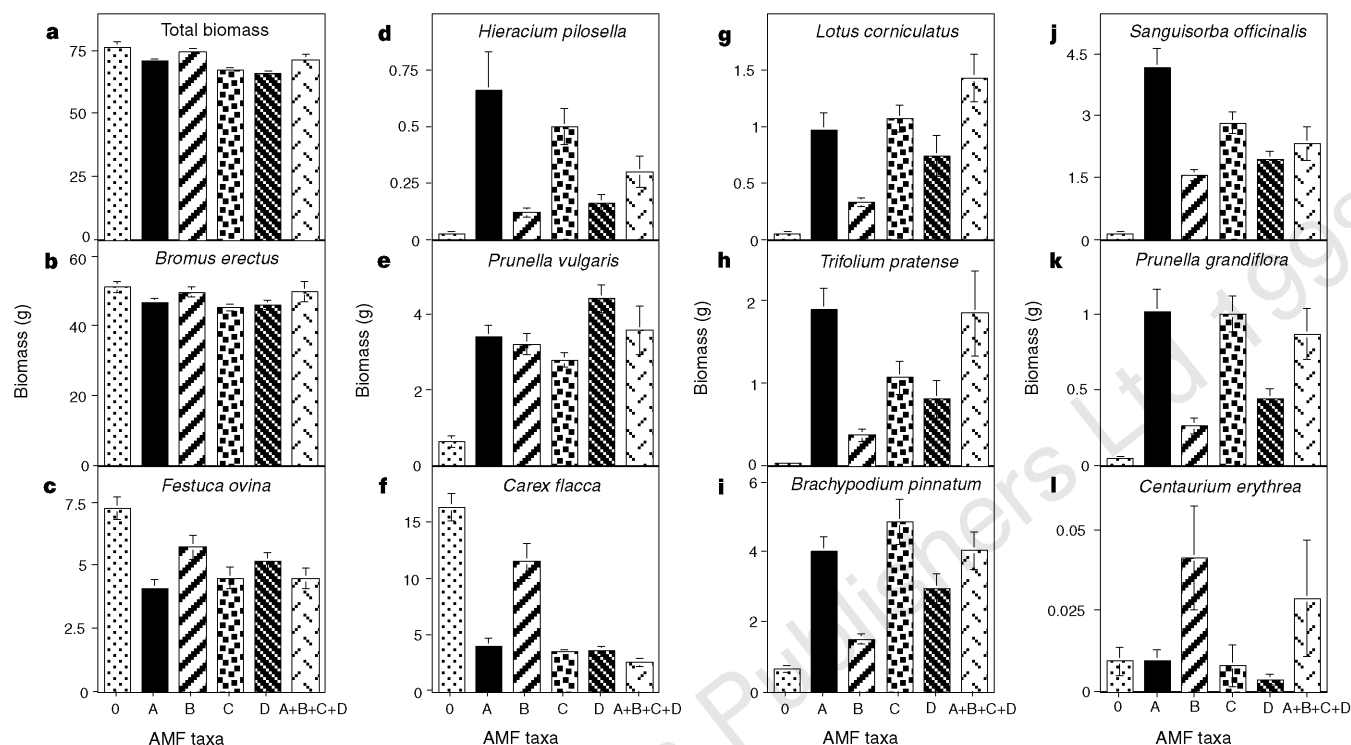


Figure 1 Above-ground biomass of individual plant species and total above-ground plant biomass (mean $\pm$ s.e.m.) in microcosms simulating calcareous grasslands, in which the composition of the AMF community was manipulated (experiment 1). Microcosms contained no AMF (0); one of four different AMF taxa, AMF A (*Glomus* sp.; Basle, Pi), AMF B (*Glomus geosporum*; BEG 18), AMF C (*Glomus* sp.; BEG 21) or AMF D (*Glomus* sp.; BEG 19); or all four AMF taxa (AMF A+B+C+D). The biomass of most individual plant species (c–k) and the total plant biomass (a) differed significantly among the six treatments (ANOVA²⁸; $P \leq 0.001$). The biomass of *Bromus erectus* (b) and of *Centaurium erythraea* (l) did not differ significantly among the treatments.

Bromus erectus, the dominant plant species in our ecosystems (Fig. 1b), the total biomass of plants in the microcosms still differed slightly, but significantly, among the AMF treatments (Fig. 1a). The biomass of several plant species in microcosms containing all four AMF taxa was roughly equal to the biomass of these plant species in those treatments that included the single AMF taxa that induced the largest growth response of the plant species (Fig. 1).

To determine whether the effects of the different AMF taxa could result in a significant alteration in the overall structure and composition of the grasslands, we used the biomass data of each of the 11 plant species as separate variables in a multivariate canonical discriminant analysis. This analysis showed that the overall structure of the plant communities varied significantly with most of the treatments with different AMF taxa (multiple analysis of variance (MANOVA) F -ratio for AMF treatment excluding the non-mycorrhizal treatment: $F_{44,97} = 5.06$, $P \leq 0.0001$).

Taken together, the results of experiment 1 show that the presence of mycorrhizal fungi is required to maintain a basic level of plant biodiversity and that, at low AMF diversity, an alteration in the composition and number of AMF taxa can lead to large fluctuations in the structure and composition of the plant community. The large changes in overall community composition were mainly manifested through a marked alteration in the biomass of subdominant species and not an alteration in the dominant grass *Bromus erectus*. The lack of a pronounced effect on *Bromus* shows that, at low AMF-species richness, changes in the community structure did not lead to marked differences in the total biomass of the microcosms. Furthermore, a reduction of AMF biodiversity from four to one AMF taxa leads to a decrease in biomass of several plant species and to a change in structure of these microcosms.

From these results, we proposed that both plant biodiversity and ecosystem productivity will increase with increasing numbers of

AMF species, because of the added beneficial effect of each single AMF species. To test this hypothesis, we set up a parallel, independent field experiment at the same time that we were running experiment 1. In the field experiment, we manipulated the number of native AMF species in 70 macrocosms simulating North American old-field ecosystems (experiment 2). To ensure an unbiased test of mycorrhizal-species richness, each replicate of each treatment received randomly selected AMF species from a pool of 23 AMF, all of which were isolated from the same site. Each macrocosm received exactly the same mixture of plant seeds from 15 plant species (see Methods for plant species list). Both the plant biodiversity, as measured by Simpson's diversity index (Fig. 2a), and productivity above and below ground (Fig. 2b, c) increased with increasing AMF-species richness. The lowest plant biodiversity and productivity were found in those plots without AMF or with only a few AMF species. In contrast, plant biodiversity and productivity were highest when eight or fourteen AMF species were present. Our results show that both the productivity and the diversity of plants in a given ecosystem can be dependent on the diversity of the fungal symbionts and, therefore, suggest that AMF should be considered as determinants of plant diversity in natural ecosystems.

Our results also indicate a mechanistic explanation for the effects of mycorrhizal-species richness on productivity and plant biodiversity. Increased AMF-species richness led to a significant increase in the length of mycorrhizal hyphae in the soil (Fig. 2d), to a decreased soil phosphorus concentration (Fig. 2e) and to an increased phosphorus content in plant material (Fig. 2f). Thus, increasing AMF biodiversity resulted in more efficient exploitation of soil phosphorus and to a better use of the resources available in the system. With a design such as that of experiment 2, increasing AMF-species richness might increase the chance of including one very effective isolate. The data from experiment 1 also indicate that

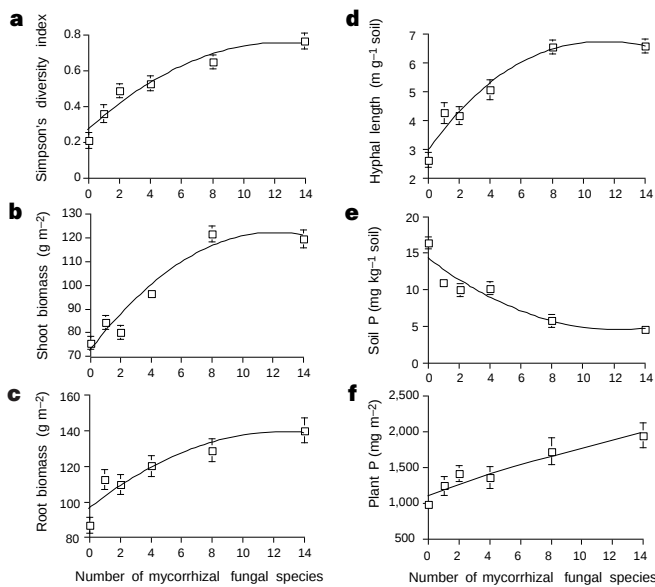


Figure 2 The effect of AMF-species richness on different parameters. Effects on: **a**, Simpson's diversity index (fitted curve is $y = 0.271 + 0.077x - 0.003x^2$; $r^2 = 0.63$; $P \leq 0.0001$); **b**, shoot biomass ($y = -0.334x^2 + 8.129x + 72.754$; $r^2 = 0.69$; $P \leq 0.0001$); **c**, root biomass ($y = -0.265x^2 + 6.772x + 96.141$; $r^2 = 0.55$; $P \leq 0.0001$); **d**, length of external mycorrhizal hyphae in soil ($y = 0.001x^3 - 0.046x^2 + 0.756x + 2.979$; $r^2 = 0.60$; $P \leq 0.0001$); **e**, soil phosphorus concentration ($y = 0.065x^2 - 1.593x + 14.252$; $r^2 = 0.67$; $P \leq 0.0001$); and **f**, total plant phosphorus content (linear relationship; $y = 61.537x + 1156.281$; $r^2 = 0.48$; $P \leq 0.001$), in macrocosms simulating North American old-field ecosystems (experiment 2). Squares represent means ($\pm$ s.e.m.).

this possibility exists, but these results also show that different plant species benefit from different AMF taxa, which could explain the effects of increasing AMF-species richness. Furthermore, independent control experiments, in which inoculum densities of effective taxa were manipulated, never resulted in AMF hyphal densities as high as those seen as a result of species-rich AMF treatments in experiment 2. The observed differential effects of AMF species on the growth of plant species (as shown in experiment 1) and the potential resulting positive feedbacks between specific plant–fungal combinations¹⁵ may, therefore, explain why an increase in the richness of AMF species (experiment 2) led to increased hyphal foraging capacity, improved resource use and increased productivity and plant biodiversity. Increasing plant biodiversity has been shown to lead to greater ecosystem productivity^{3,4}. The significant positive correlation between a plant–biodiversity index and ecosystem productivity in experiment 2 (Pearson's $r = 0.70$, $P \leq 0.001$) shows that such a relationship exists. We have shown here that this relationship occurs as a result of altering the richness of AMF species.

Our results emphasize the importance of the mycorrhizal symbiosis as a determinant of plant biodiversity, ecosystem variability and productivity. The loss of AMF biodiversity, which occurs in agricultural systems^{20,23} could, therefore, decrease both plant biodiversity and ecosystem productivity while increasing ecosystem instability. The loss of biodiversity in soils represents an understudied field of research which requires more attention. The present reduction in biodiversity on Earth and its potential threat to ecosystem stability and sustainability^{1,4} can only be reversed or stopped if whole ecosystems, including ecosystem components other than plants, are protected and conserved. □

Methods

Experiment 1. We established 48 microcosms simulating European calcareous grassland under sterile greenhouse conditions in containers measuring

$26.5 \times 17 \times 18 \text{ cm}^3$. These containers were filled with a 7.38-kg soil mixture of autoclaved quartz sand and autoclaved calcareous grassland soil (1:1 v/v). The microcosms were inoculated with 100-g soil inoculum of one AMF species (four single AMF-species treatments) or a mixture of the four AMF species, or were inoculated with an autoclaved (121 °C; 30 min) soil mixture of these four AMF species (the non-mycorrhizal control). The experiment was set up as a randomized block design in which each treatment was replicated eight times and each block was randomized every 2 weeks. In each microcosm, 70 seedlings of 11 calcareous grassland plant species were planted at random and put at fixed distances from each other. The number of seedlings of each plant species corresponded to natural abundances in a calcareous grassland in Switzerland. Microcosms received 380 ml filtered washing of soil inoculum (without AMF propagules) to correct for possible differences in microbial communities²⁴.

The four different AMF species, all belonging to the genus of *Glomus* (class: Zygomycete), were isolated from a calcareous grassland in Switzerland^{21,22}, and the calcareous grassland soil was obtained from a similar site. Detailed information about the four AMF taxa is published elsewhere^{21,22} and further information on BEG AMF taxa can be obtained at <http://www.bio.uk.ac.uk/cinetpubwwwroot/beg/begdatabase/amfreq%2Dreg.htm>. All AMF isolates are both morphologically and genetically different from each other, but as three of the four taxa are undescribed we refer to them here as taxa and not species.

Microcosms were maintained in the greenhouse during two growing seasons and received a winter period from January to March 1997. Microcosms were watered three times a week and each microcosm was adjusted to an equal soil water content every two weeks by weighing. Above-ground plant parts were cut per plant individual in four regular harvests (mowing level 2.5 cm above soil surface) and were cut to the soil surface at the final harvest in December 1997. Total above-ground biomass of each plant species was determined, after drying at 70 °C, by adding the biomass of each of the five harvests.

Experiment 2. We established 70 macrocosms simulating North American old-field ecosystems under field conditions and inoculated them with 1 kg AMF inoculum containing 1, 2, 4, 8 or 14 species, which were randomly selected from a pool of 23 different AMF species. All AMF species were isolated from the Long-Term Mycorrhizae Research Site of the University of Guelph. The species were *Acaulospora denticulata*, *Acaulospora morrowiae*, *Acaulospora spinosa*, *Acaulospora* spp. 1 and 2 (undescribed), *Entrophospora colombiana*, *Gigaspora gigantea*, *Gigaspora margarita*, *Gigaspora rosea*, *Glomus claroideum*, *Glomus etunicatum*, *Glomus intraradices*, *Glomus macrocarpum*, *Glomus mosseae*, *Glomus* spp. 1–4 (undescribed), *Scutellospora calospora*, *Scutellospora dipurpurens*, *Scutellospora heterogama* and *Scutellospora pellucida*. All mycorrhizal taxa were morphologically distinct from each other and are referred to here as species. The non-mycorrhizal control received 1 kg sterilized (γ -ray-irradiated) AMF inoculum. All macrocosms received 1 litre filtered washing of AMF inoculum to correct for possible differences in microbial communities²⁴. We set up the experiment in a randomized design, each treatment being replicated ten times.

Each macrocosm was showered with a seed rain consisting of 100 seeds from each of the following 15 most abundant plant species of the research site: *Agrostis gigantea*, *Bromus inermis*, *Poa compressa*, *Achillea millefolium*, *Aster cordifolius*, *Aster novae-angliae*, *Chrysanthemum leucanthemum*, *Daucus carota*, *Euthamia graminifolia*, *Fragaria virginiana*, *Plantago lanceolata*, *Ranunculus acris*, *Rudbeckia hirta*, *Geum macrophyllum* and *Solidago canadensis*.

Each macrocosm, $1 \times 0.75 \times 0.25 \text{ m}^3$ contained a 90-kg mixture of γ -ray irradiated sand and soil (1:1 v/v). We fertilized macrocosms each month with a half-strength, low-phosphorus, Hoagland nutrient solution. After one growing season, plants were harvested and root and shoot biomass were determined after drying for 48 h at 60 °C. Plant biodiversity was assessed using Simpson's diversity index²⁵ on individual species shoot-biomass data. Plant biomass was assessed as total shoot and root biomass for the entire plant community. Shoot plant phosphorus concentrations were determined for each plant species and summed to give total shoot phosphorus content for each entire macrocosm. Root phosphorus concentration was measured from a mixed sample of roots, and total root phosphorus content was estimated for each entire macrocosm using the root biomass data. Soil phosphorus was assessed as bicarbonate-extractable soil phosphorus²⁶. We measured AMF hyphal length on membrane filters²⁷.

In the non-mycorrhizal control, there was less than 3 m g^{-1} non-mycorrhizal hyphae; this represents a background count of non-mycorrhizal fungi in each treatment. In the non-mycorrhizal control treatments of experiments 1 and 2, plant roots were not colonized by AMF and no spores of AMF were found. For further information, readers may contact J.N.K.

Received 28 July; accepted 10 September 1998.

- Schulze, E. D. & Mooney, H. A. (eds) *Biodiversity and Ecosystem Function* (Springer, Berlin, 1993).
- Tilman, D. & Downing, J. A. Biodiversity and stability in grasslands. *Nature* **367**, 363–365 (1994).
- Naem, S., Thompson, L. J., Lawler, S. P., Lawton, J. H. & Woodfin, R. M. Declining biodiversity can alter the performance of ecosystems. *Nature* **368**, 734–737 (1994).
- Tilman, D., Wedin, D. & Knops, J. Productivity and sustainability influenced by biodiversity in grassland ecosystems. *Nature* **379**, 718–720 (1996).
- Hooper, D. U. & Vitousek, P. M. The effects of plant composition and diversity on ecosystem processes. *Science* **277**, 1302–1305 (1997).
- Aarssen, W. L. Ecological combining ability and competitive combining ability in plants: towards a general evolutionary theory of coexistence in systems of competition. *Am. Nat.* **122**, 707–731 (1983).
- Grace, J. D. & Tilman, D. (eds) *Perspectives on Plant Competition* (Academic, San Diego, 1990).
- Ricklefs, R. W. Environmental heterogeneity and plant species diversity: a hypothesis. *Am. Nat.* **111**, 376–381 (1977).
- Tilman, D. *Resource Competition and Community Structure* (Princeton Univ. Press, Princeton, 1982).
- Grubb, P. The maintenance of species richness in plant communities: the importance of the regeneration niche. *Biol. Rev.* **52**, 107–145 (1977).
- Huston, M. A. General hypothesis of species diversity. *Am. Nat.* **113**, 81–101 (1979).
- Brown, V. K. & Gange, A. C. Herbivory by soil dwelling insects depresses plant species richness. *Funct. Ecol.* **3**, 667–671 (1989).
- Dobson, A. & Crawley, M. J. Pathogens and the structure of plant communities. *TREE* **9**, 393–398 (1994).
- de Ruiter, P. C., Neutel, A. & Moore, J. C. Energetics, patterns of interaction strengths and stability in real ecosystems. *Science* **269**, 1257–1260 (1995).
- Bever, J. D., Westover, K. M. & Antonovics, J. Incorporating the soil community into plant population dynamics: the utility of the feedback approach. *J. Ecol.* **85**, 561–571 (1997).
- Smith, S. E. & Read, D. J. *Mycorrhizal Symbiosis* 2nd edn (Academic, San Diego, 1997).
- Grime, J. P., Mackey, J. M. L., Hillier, S. H. & Read, D. J. Floristic diversity in a model system using experimental microcosms. *Nature* **328**, 420–422 (1987).
- Walker, C., Mizue, C. W. & McNabb, H. S. Populations of endogoneaceous fungi at two locations in central Iowa. *Can. J. Bot.* **60**, 2518–2529 (1982).
- Johnson, N. C., Tilman, D. & Wedin, D. Plant and soil controls on mycorrhizal fungal communities. *Ecology* **73**, 2034–2042 (1992).
- Helgason, T., Daniell, T. J., Husband, R., Fitter, A. H. & Young, J. P. Y. Ploughing up the wood-wide web? *Nature* **394**, 431 (1998).
- Streitwolf-Engel, R., Boller, T., Wiemken, A. & Sanders, I. R. Clonal growth traits of two *Prunella* species are determined by co-occurring arbuscular mycorrhizal fungi from a calcareous grassland. *J. Ecol.* **85**, 181–191 (1997).
- van der Heijden, M. G. A., Boller, T., Wiemken, A. & Sanders, I. R. Different arbuscular mycorrhizal fungal species are potential determinants of plant community structure. *Ecology* **79**, 2082–2091 (1998).
- Johnson, N. C. Can fertilisation of soil select less mutualistic mycorrhizae? *Ecol. Appl.* **3**, 749–757 (1993).
- Koide, R. T. & Li, M. Appropriate controls for vesicular arbuscular mycorrhizal research. *New Phytol.* **111**, 35–44 (1989).
- Krebs, C. J. *Ecological Methodology* (Harper Collins Publishers, New York, 1989).
- Olsen, S. R. & Sommers, L. W. In *Methods of Soil Analysis* part 2 403–430 (ed. A. L. Page) (Am. Soc. Agron., Madison, 1982).
- Jakobsen, I. & Roser, L. Carbon flow into soil and external hyphae from roots of mycorrhizal cucumber plants. *New Phytol.* **115**, 77–83 (1990).
- SAS/STAT user's guide. Release 6.04 (eds Allen, J. F., Ashton, J. J., Cohen, B. L., Cornell, R. Jr, Early, R., et al.) (SAS Institute, Cary, NC, 1989).

Acknowledgements. This work was supported by the Swiss National Science Foundation and the Natural Sciences and Engineering Research Council of Canada. We thank P. Leadley, D. Read, P. Salzer, B. Schmid, J. Stocklin and D. Tilman for comments.

Correspondence and requests for materials should be addressed to I.R.S. (e-mail: sanders@ubacu.unibas.ch).

Attention improves or impairs visual performance by enhancing spatial resolution

Yaffa Yeshurun & Marisa Carrasco

Department of Psychology, New York University, New York, New York 10003, USA

Covert attention, the selective processing of visual information at a given location in the absence of eye movements, improves performance in several tasks, such as visual search and detection of luminance and vernier targets^{1–6}. An important unsettled issue is whether this improvement is due to a reduction in noise (internal or external)^{6–9}, a change in decisional criteria^{10,11}, or signal enhancement^{3,5,12}. Here we show that attention can affect

performance by signal enhancement. For a texture segregation task in which performance is actually diminished when spatial resolution is too high, we observed that attention improved performance at peripheral locations where spatial resolution was too low, but impaired performance at central locations where spatial resolution was too high^{4–12}. The counterintuitive impairment of performance that we found at the central retinal locations appears to have only one possible explanation: attention enhances spatial resolution.

We previously demonstrated that when a spatial cue directs covert attention to an upcoming target location, observers' performance improves for stimuli designed to measure spatial resolution (for example, the Landolt-square—a square with a small gap on one side)⁵. This 'peripheral cue' putatively draws covert attention to its location automatically^{2,3,13}. Because the display characteristics ensured that neither a reduction in noise nor a change in decisional criteria could explain this attentional facilitation, this finding indicated that attention could enhance spatial resolution at the cue location⁵. Similarly, in visual search tasks in which observers'

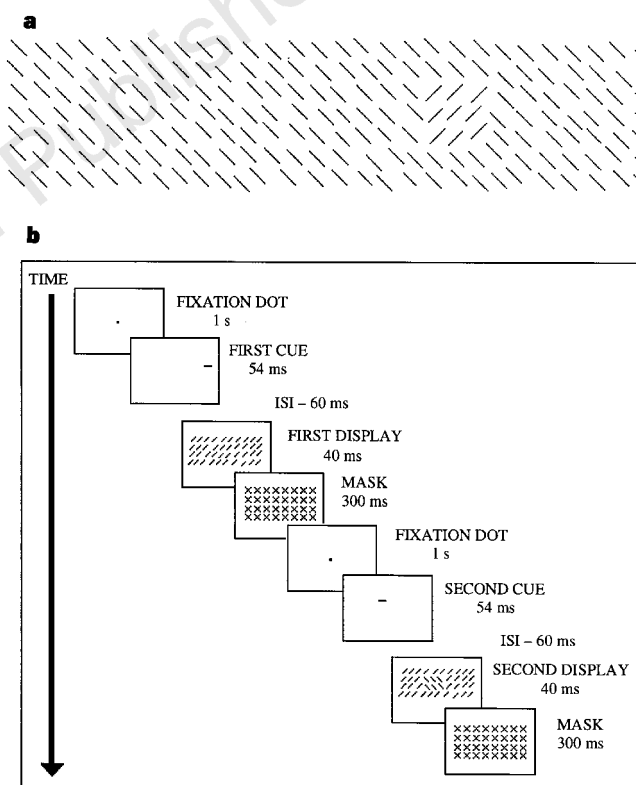


Figure 1 Texture segregation task. **a**, The display consisted of a 2×1.5 cm target-texture, composed of 3×3 lines (oriented at 45° or 135°), embedded in a background-texture composed of 287 lines (7 rows $\times$ 41 columns, subtending 5×28 cm) whose orientation was orthogonal to the target. The elements were jittered by 0.3 cm. From a viewing distance of 57 cm, the target subtended $2 \times 1.5^\circ$ of visual angle and the texture display subtended $5 \times 14^\circ$ of visual angle to each side of the centre of the display. The target appeared equally often in each interval (50% of the time) and was centred at any of 35 possible locations along the horizontal meridian in a random order. **b**, Each interval began with a fixation dot at the centre, followed by a brief cue. The cue was either 'peripheral' (a green horizontal bar of 0.3×0.6 cm appearing 0.3 cm above the target location) or neutral (two horizontal lines of 0.3×28 cm appearing above and below the display). After an interstimulus interval (ISI), the texture was displayed for an average of 40 ms. The ISI was set individually to keep overall performance level 75% correct; display duration ranged from 15 to 50 ms. A mask, with crosses as elements, followed the stimulus. Observers were asked to indicate the interval containing the target by pressing one of two keys. The order of the 100 practice trials, as well as that of the 288 experimental trials, was randomized.

performance is slower and less accurate as target eccentricity increases, due to the lower spatial resolution of the periphery¹⁴, cueing the target location diminished this eccentricity effect¹. In another study⁶, when observers searched for a slightly tilted target, cueing the target location decreased the detrimental effect caused by vertical distracters; the authors suggested that focal attention changes the size of the 'stimulus analyser', permitting a finer-scale analysis of a target⁶.

We present a critical test of the 'resolution hypothesis', which predicts that attention can actually enhance spatial resolution, so that we can resolve finer details at the attended location. We explored the effects of spatial attention on a task in which performance would be diminished by heightened resolution. If attention indeed enhances resolution, performance at the attended location should be impaired, not improved. Such impairment could not be predicted by previous models of visual attention^{7–11,15,16}. The task involves the detection of a texture target of one orientation appearing at a large range of eccentricities in a background of an orthogonal orientation (Fig. 1a). Unlike in most visual tasks^{1,5,9,14,17} observers' performance does not peak when the target appears at foveal locations, where resolution is highest^{9,17}, but rather at mid-peripheral locations, and drops towards central or farther peripheral locations^{18–21}. Psychophysical and physiological evidence indicates that we process visual stimuli by means of parallel spatial filters, each tuned to a band of spatial frequencies^{9,17}. The central performance drop may be due to the fovea being more sensitive to high spatial frequencies for which neural processing is slower²⁰. Moreover, this drop could reflect a mismatch between the average size of spatial filters at the fovea and the scale of the texture. Spatial filters at the fovea may be too small for the scale of the texture¹⁸, such that the foveal resolution may be too high for this texture. As retinal eccentricity increases, the average size of the filters becomes larger and their resolution decreases^{9,17}; filter size is presumably optimal around the peak of performance. At farther eccentricities, the filters may be too big and their lower resolution would limit performance¹⁸.

We hypothesized that if attention enhances spatial resolution, attending to the target location will improve performance where the resolution is too low (periphery) but will impair performance where resolution is already too high for the task (central locations). We employed a two-interval forced-choice task. Observers indicated which of two intervals contained a texture-target whose orientation

differed from that of the background (Fig. 1a). Half the trials were 'cued trials' in which a peripheral cue indicated the target location in the interval containing the texture target, and a non-target location in the interval without a target. The cue was a 100% valid cue because the target could only appear at the cued location; however, it did not signal which interval was more likely to contain the target because both displays were preceded by a cue¹. The rest were 'neutral trials' in which a pair of lines in both intervals indicated that the target had an equal probability of appearing at any location. The viewing distance was 57 cm for all 28 naive observers. Brief presentation times precluded eye movements between the onset of the cue and the offset of the texture display (Fig. 1b).

In neutral trials, performance peaked when the target appeared at 5° eccentricity, not at the fovea, replicating the central performance drop found in a similar texture task^{18–21}. Cueing condition and target eccentricity interacted significantly (Fig. 2a), in a manner consistent with the resolution hypothesis. Accuracy was higher for the cued than the neutral trials at all target eccentricities except at foveal locations (<1°), where it was lower. These results are inconsistent with previous models of attention which predict that attention always helps and never hinders performance, but they are consistent with our hypothesis. Given that performance is worse at the fovea because its spatial filters are too small and have too high a resolution for the scale of the texture^{18,21}, further increasing resolution at foveal locations led to a more pronounced drop in performance. Such impairment could result from an attentional mechanism that enhances resolution by effectively decreasing the average size of filters at the attended location (see below). The same mechanism should improve performance at regions where spatial filters are too large and have too low a resolution for the scale of the texture.

Performance peaks at different eccentricities when the scale of the texture is manipulated by changing the size of the textural elements, the space between them, or the viewing distance^{18–20}. Enlarging the scale of the texture shifts the peak of performance to farther eccentricities, supporting the idea that segregating larger textures requires larger filters, more abundant at farther eccentricities¹⁸. If by increasing the scale of the texture, its mismatch with the size of the filters extends farther towards the periphery, and attending to a location is similar to reducing the size of spatial filters, then cueing should impair performance for a wider range of eccentricities. In experiment 2, we tested this hypothesis by doubling the scale of the

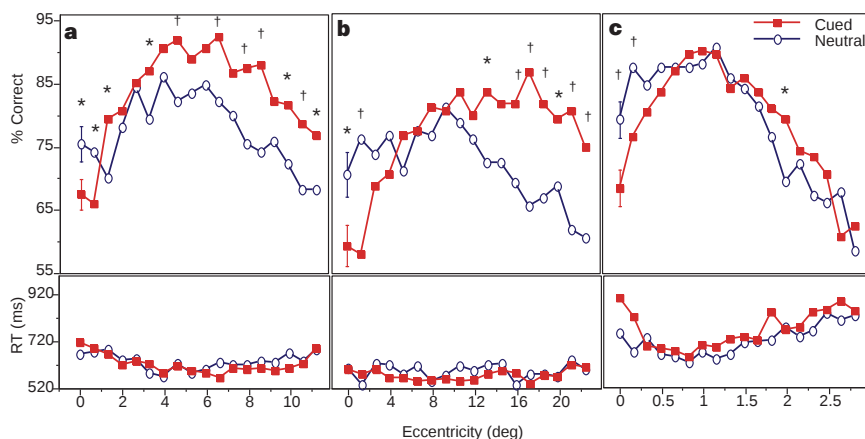


Figure 2 Observers' performance as a function of cueing condition and target eccentricity. **a**, Experiment 1; **b**, experiment 2; and **c**, experiment 3 (viewing distance of 57 cm, 28 cm, and 228 cm, respectively). Within-observers ANOVAs demonstrated significant interactions ($P < 0.001$); accuracy was higher for the cued (squares) than the neutral (circles) trials at more peripheral eccentricities but was lower at central locations (exp. 1: 0–1°; exp. 2: 0–5°; exp. 3: 0–0.66°). Analysis of reaction time (RT) indicated that neither the main effects of cueing and eccentricity nor their interaction were significant (bottom panels). Thus, there

were no speed-accuracy tradeoffs. In experiment 2, halving the viewing distance doubled the target eccentricity range; in experiment 3, increasing the viewing distance by a factor of 4 reduced the range accordingly (see abscissa scales). Note the shift in performance peak and the range in which performance was impaired by the cue. Error bars correspond to the average ± 1 s.e. for each condition; symbols denote statistically significant differences at $P < 0.05$ (asterisks) or $P < 0.01$ (daggers) according to least significant differences (LSD) post hoc comparisons.

texture; the viewing distance for 20 naive observers was reduced to 28 cm.

Consistent with previous studies^{18–20}, doubling the scale shifted the peak of performance to a farther eccentricity (from 5 to 7.6°). Again, we observed a significant interaction of cueing and eccentricity (Fig. 2b). Performance in the cued trials improved in peripheral regions but diminished at the four central locations, which extend into the parafovea. Moreover, with the larger texture scale, the cue impaired performance for a larger range of eccentricities (0–5°) than in experiment 1. In experiment 3, in contrast, the scale of the texture was reduced for 23 naive observers who viewed the display from a distance of 228 cm, and the range of impaired performance shrank to only 0–0.66° (Fig. 2c). These results suggest that the ranges of attentional impairment and enhancement are mediated by the scale of the texture. Specifically, the finding that target detection at a given eccentricity was either impaired or improved by the cue, depending only on the scale of the texture, rules out the possibility that attention is unable to enhance processing of foveal and parafoveal stimuli in this task. Rather, this impairment results from the characteristics of a given texture and the effects of covert attention.

The mere fact that attentional deployment affected performance is noteworthy. Most models of visual processing portray texture segregation based on orientation differences as an early pre-attentive task²², in which information (for example, simple features) is extracted from the display effortlessly and in parallel. Such models cannot account for the effect found here. This effect is consistent, however, with psychophysical studies showing that attention can improve performance in tasks considered to be pre-attentive, including feature search^{1,6}, feature discrimination²³, spatial acuity⁵ and orientation acuity⁶ tasks. This effect also supports the finding that performance in a simple feature search task declines when additional attentional demands are imposed^{24,25}. Similarly, neurophysiological studies demonstrate that attention can modify the responses of early visual areas such as V1, V2, V4 and MT/MST neurons^{26–30}. This body of psychophysical and neurophysiological evidence indicates that attentional mechanisms can influence early vision.

Either decreasing the viewing distance (Fig. 3a) or directing attention to the target location (Fig. 3b) similarly shifted the peak of performance to farther eccentricities, suggesting that both

manipulations increased the mismatch between texture scale and the size of the filters. The former may have done so by increasing the scale of the texture and the latter by decreasing the size of the filters responsible for texture segregation. Like moving closer to the display, attention allows us to resolve better the various details in front of us, which would almost always be advantageous. Yet when a more global inspection of the display is required, for example when one is appreciating an impressionist painting, moving closer is not the optimal strategy.

Physiological studies support the hypothesis that attention can enhance spatial resolution by reducing the size of spatial filters at the attended area. When attended and unattended stimuli are both within a cell's receptive field, the neuronal response is primarily determined by the attended stimuli; responses to the unattended stimuli are attenuated^{26,27}, as if the cell's receptive field shrinks around the attended stimulus^{26–28}. These authors proposed that attentional modulation of sensory processing is accomplished in two stages: first, top-down signals bias activity in favour of the neurons representing the relevant location, and then these favoured neurons compete with other neurons, ultimately suppressing their response. This competition may result from mutual inhibition between cells or between the inputs to the cells^{27,28}, and its outcome could effectively reduce the cell's receptive field, allowing finer spatial resolution. Alternatively, enhanced resolution at the attended location could result from increased sensitivity of the neurons with the smallest receptive fields at the attended area^{4,5}, which in turn may inhibit neurons with larger receptive fields there¹⁷. According to this hypothesis, the overall sensitivity of the attended region would shift towards higher spatial frequencies. However, the possibility that the central performance drop is mediated by interference between high- and low-frequency information is controversial^{18,21}.

We conclude that attentional facilitation in visual tasks may reflect a combination of mechanisms—such as noise reduction, decisional factors and signal enhancement^{5,6,10,12}. This study supports the last mechanism because directing attention improved performance in peripheral locations, but impaired it at foveal and parafoveal locations. These findings imply that attention increased the mismatch between the texture scale and the size of central spatial filters, and provide strong support for the hypothesis that attention can enhance spatial resolution. □

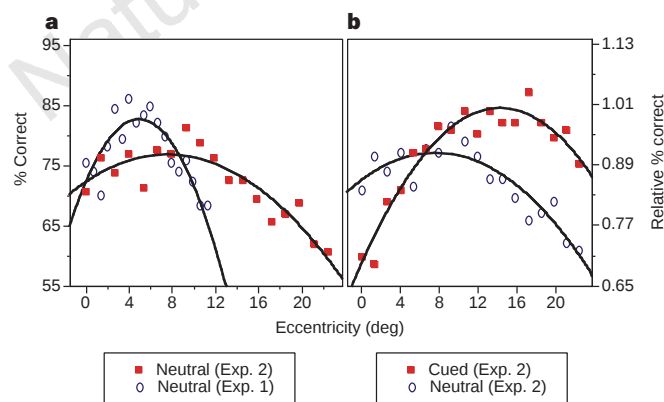


Figure 3 Observers' performance as a function of viewing distance and cueing condition. For both panels, performance is depicted as a function of per cent correct (left axis) and relative per cent correct, with respect to optimal performance (right axis). The data from experiments 1 and 2 were fitted to second-order polynomials. The fits' R^2 ranged from 0.8 to 0.97. **a**, Comparing the neutral trials of the two experiments demonstrates the effect of viewing distance on performance. Viewing the display from half the distance shifted the performance peak from 5° to 7.6°. **b**, Comparing the cued and neutral trials of experiment 2 demonstrates the effect of cueing on performance. Directing attention shifted the performance peak from 7.6° to 13.3°.

Received 17 August; accepted 14 September 1998.

- Carrasco, M. & Yeshurun, Y. The contribution of covert attention to the set-size and eccentricity effects in visual search. *J. Exp. Psychol. Hum. Percept. Perform.* **24**, 673–692 (1998).
- Nakayama, K. & Mackeben, M. Sustained and transient components of focal visual attention. *Vision Res.* **29**, 1631–1646 (1989).
- Posner, M. I. Orienting of attention. *Q. J. Exp. Psychol.* **32**, 3–25 (1980).
- Balz, G. W. & Hock, H. S. The effect of attentional spread on spatial resolution. *Vision Res.* **37**, 1499–1510 (1997).
- Yeshurun, Y. & Carrasco, M. Spatial attention improves performance in spatial resolution tasks. *Vision Res.* **39**, 293–305 (1999).
- Morgan, M. J., Ward, R. M. & Castet, E. Visual search for a tilted target: tests of spatial uncertainty models. *Q. J. Exp. Psychol.* **51A**, 343–370 (1998).
- Palmer, J. Set-size effects in visual search: the effect of attention is independent of the stimulus for simple tasks. *Vision Res.* **34**, 1703–1721 (1994).
- Sperling, G. & Doshier, B. A. in *Handbook of Perception and Human Performance* (eds Boff, K. R., Kaufman, L. & Thomas, J. P.) Vol. 1, Ch. 2, 1–65 (Wiley, New York, 1986).
- Graham, N. *Visual Pattern Analyzers* (Oxford University Press, NY, 1989).
- Kinchla, R. A. Attention. *Annu. Rev. Psychol.* **43**, 711–742 (1992).
- Shaw, M. L. in *Attention and Performance X* (eds Bouma, H. & Bouwhuis, D. G.) 109–120 (Erlbaum, Hillsdale, NJ, 1984).
- Lu, Z. L. & Doshier, B. A. External noise distinguishes attention mechanisms. *Vision Res.* **38**, 1183–1198 (1998).
- Jonides, J. in *Attention and Performance IX* (eds Long, J. B. & Baddeley, A. D.) 187–204 (Erlbaum, Hillsdale, NJ, 1981).
- Carrasco, M. & Frieder, K. S. Cortical magnification neutralizes the eccentricity effect in visual search. *Visual Res.* **37**, 63–82 (1997).
- Duncan, J. & Humphreys, G. W. Beyond the search surface: visual search and attentional engagement. *J. Exp. Psychol. Hum. Percept. Perform.* **18**, 578–588 (1992).
- Wolfe, J. M. Guided search 2.0: a revised model of visual search. *Psych. Bull. Rev.* **1**, 202–238 (1994).
- DeValois, R. L. & DeValois, K. K. *Spatial Vision* (Oxford University, New York, 1988).
- Gurnsey, R., Pearson, P. & Day, D. Texture segmentation along the horizontal meridian: nonmonotonic changes in performance with eccentricity. *J. Exp. Psychol. Hum. Percept. Perform.* **22**, 738–757 (1996).
- Joffe, K. M. & Scialfa, C. T. Texture segmentation as a function of eccentricity, spatial frequency and target size. *Spatial Vis.* **9**, 325–342 (1995).

20. Kehr, L. Central performance drop on perceptual segregation tasks. *Spatial Vis.* **4**, 45–62 (1989).
21. Morikawa, K. Peripheral advantage in texture segmentation: the role of spatial and temporal factors. *Invest. Ophthalmol. Vis. Sci.* **38** (suppl.), 2986 (1997).
22. Julesz, B. Textons, the elements of texture perception, and their interactions. *Nature* **290**, 91–97 (1981).
23. Cheal, M. L. & Lyon, D. R. Benefits from attention depend on the target type in location-precued discrimination. *Acta Psychol.* **81**, 243–267 (1992).
24. Joseph, J. S., Chun, M. M. & Nakayama, K. Attentional requirements in a 'preattentive' feature search task. *Nature* **387**, 805–807 (1997).
25. Lee, D. K., Koch, C. & Braun, J. Spatial vision thresholds in the near absence of attention. *Vision Res.* **37**, 2409–2418 (1997).
26. Moran, J. & Desimone, R. Selective attention gates visual processing in the extrastriate cortex. *Science* **229**, 782–784 (1985).
27. Luck, S. J., Chelazzi, L., Hillyard, S. A. & Desimone, R. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J. Neurophysiol.* **77**, 24–42 (1997).
28. Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* **18**, 193–222 (1995).
29. Motter, B. M. Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *J. Neurophysiol.* **70**, 909–919 (1993).
30. Treue, S. & Maunsell, J. H. R. Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature* **382**, 539–541 (1996).

Acknowledgements. This study was supported by an NSF National Young Investigator Grant to M.C. and a Katzell Summer Fellowship to Y.Y. We thank K. Adolph, L. Cameron, J. Fernández, K. Frieder, P. Glimcher, M. Landy, L. Maloney, B. McElree, D. Pelli, E. Phelps and S. Wolfson for their comments on a draft of this manuscript.

Correspondence and requests for materials should be addressed to M.C. (e-mail: marisa@psych.nyu.edu).

Hippocampal lesions disrupt navigation based on cognitive maps but not heading vectors

John M. Pearce, Amanda D. L. Roberts & Mark Good

School of Psychology, Cardiff University, Cardiff CF1 3YG, UK

Animals can find a hidden goal in several ways. They might use a cognitive map that encodes information about the geometric relationship between the goal and two or more landmarks¹. Alternatively, they might use a heading vector that specifies the direction and distance of the goal from a single landmark². Rats with damage to the hippocampus have difficulty in finding a hidden goal³. Here we determine which of the above strategies is affected by such damage. Rats were required to swim in a water maze to a submerged platform, which was always at the same distance and direction from a landmark. The platform and landmark remained in the same place for the four trials of each session, but they were moved to a new position at the start of a

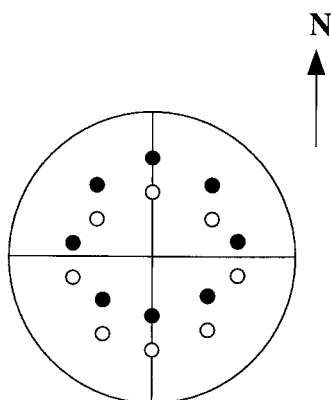


Figure 1 The eight possible positions that were occupied by the landmark (black circles) and the platform (open circles) for rats trained with the platform due south of the landmark. The diameter of the pool was 2 m, of the landmark 13 cm, and of the platform 10 cm. For half of the rats in each group of each experiment the platform was south of the landmark, and for the remainder it was north of the landmark.

session⁴. Rats with damage to the hippocampus found the platform more efficiently than did normal rats in the first trial of a session but, in contrast to normal rats, their performance did not improve during a session. Our results indicate that hippocampally damaged rats are able to navigate by means of heading vectors but not cognitive maps.

Eight male Lister-hooded rats received ibotenic-acid-induced lesions of the hippocampus⁵ and eight were subjected to a control operation. Following a period of recovery, all rats were trained to find a submerged platform in a Morris pool⁶. The platform was always 20 cm in a constant direction from a spherical black landmark on the surface of the water (Fig. 1). On completion of behavioural testing, the animals were killed and their brains prepared for histology using a standard nissle stain (Fig. 2). Figure 3 shows the mean latency to escape from the pool (to the platform) for both groups, for the first and last trial of each session of training. The performance of the control group improved across the training sessions and, within each session, performance in the last trial was superior to that in the first trial. The escape latencies of hippocampally damaged rats also decreased as training progressed, but there was little evidence of a within-session improvement in performance. In the first trial of each of the intermediate sessions of the experiment, rats with hippocampal damage found the platform more rapidly than did control rats, whereas in the last trial of these sessions the performance of the control group was superior to that of the hippocampally damaged group. Analyses of variance (ANOVAs) (evaluated at $P < 0.05$), using individual mean escape

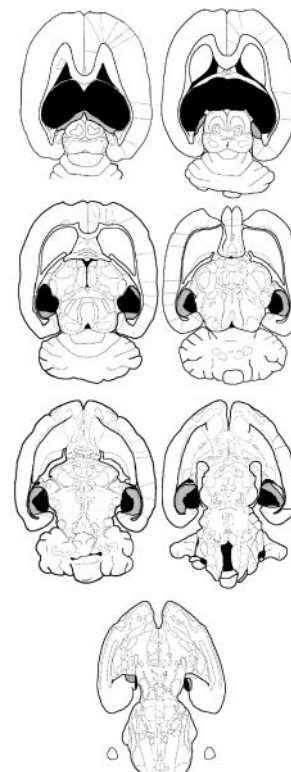


Figure 2 The maximum (dotted region) and minimum (black region) extent of the ibotenic acid lesion recreated on horizontal sections taken throughout the dorsoventral extent of the hippocampus. All hippocampally lesioned animals suffered >90% cell loss in the dorsal aspects of the hippocampus, with two animals showing minor incursion of the lesion into the subicular complex. Cell loss in the mid-dorsal and ventral aspects of the hippocampus was more variable (40–70%). Performance of the lesioned animals did not vary with the extent of hippocampal cell loss in the ventral aspects of the hippocampus. The sections are arranged in descending order through the brain from dorsal (top left) to ventral (bottom).

latencies for the 11 sessions combined, indicated that escape latencies were significantly longer in the first trial than in the fourth trial for the control group ($F = 68.55$, degrees of freedom (d.f.) = 1/14), but not for the hippocampal group ($F < 1$), and that there was a significant difference between the groups in the first trial and in the fourth trial (F values > 17.20 , d.f. = 1/28).

There are two reasons for believing that the control group identified the position of the platform by means of a cognitive map based on the cues outside the pool, rather than by using the landmark within the pool. First, the acquisition of such a map in one session would encourage the rats to search for the platform in the wrong location at the start of the next sessions, which would explain the relatively long escape latencies for the control group in the first trial of many sessions. Second, the development of such a map would explain the marked improvement in performance that was seen within a session by the control group. In contrast, there is no evidence that rats with damage to the hippocampus identified the position of the platform by means of a cognitive map based on cues outside the pool. In the absence of such a map, these rats would not search in the wrong part of the pool at the outset of a session, which would explain why their escape latencies were significantly shorter than for the control group in trial 1 of a session. At the same time, the lack of a cognitive map would prevent hippocampal rats from showing the within-session improvement that was so marked in the control group.

We studied the results further to confirm that, in the first trial of each session, during the intermediate sessions of training, control but not hippocampally damaged rats explored the region of the pool in which the platform had been placed for the previous session. Figure 4 shows the paths from the point of release to the platform taken by representative control rats and hippocampally damaged rats. Both control rats showed a tendency to explore the area in which the platform had been located in the previous session, whereas the hippocampal rats took a more direct route to the platform. To analyse these results quantitatively, we recorded the time spent by each rat in a square with an area of 0.4 m^2 that was centred on the position at which the platform had been located in the previous session. We expressed this time as a percentage of the total time spent by the rat in the pool. The mean of these

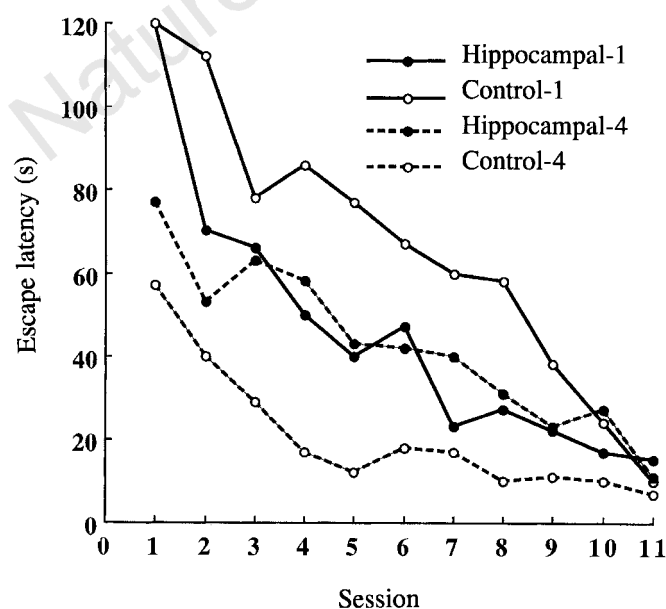


Figure 3 The mean escape latency for the first (Hippocampal-1 and Control-1) and fourth (Hippocampal-4 and Control-4) trials of each of the 11 sessions of the first experiment for the group of rats with hippocampal lesions and for the group of control rats.

percentages for the two sessions combined was 15% for the control group and 8% for the hippocampal group, which indicates that control rats spent a greater proportion of a trial in the region in which the platform had been situated previously than did hippocampal rats. An analysis of individual mean percentages indicated that this difference between the groups was significant ($t = 3.3$, d.f. = 14).

The improvement shown by the two groups in the first trial of successive sessions indicates that both hippocampally damaged and control rats were able to use the landmark as a point of reference for finding the platform. One strategy that the rats may have adopted could have been to swim to the landmark and then search at random until the platform was discovered. Alternatively, rats may have learned that the platform was in a given direction from the landmark. To differentiate between these possibilities, we conducted a test trial immediately after the final training session. In this test trial, the landmark was placed in the centre of the pool for the first time. For half of the rats in each group the platform was located at the usual direction and distance from the landmark, and for the remaining rats the platform was located at the same distance from the landmark but on the opposite side to normal. When the platform was located on the normal side of the landmark, both groups found it relatively rapidly. The mean escape latency in this test trial was 9.8 s for the hippocampally damaged rats and 10.2 s for the control rats. In contrast, both groups experienced difficulty in finding the platform when it was on the opposite side of the landmark to normal. The mean escape latency for rats tested in this way was 59.4 s for the hippocampally damaged rats and 54.6 s for the control rats. A two-way ANOVA showed that there was a significant effect of moving the platform to the opposite side to normal ($F = 11.1$, d.f. = 1/12), but the effect of group, and the interaction between these effects were not significant (F values < 1). These results show that both groups appreciated the direction in which the platform could be found with respect to the landmark. We cannot specify the source of this directional information, but it could be based on an internal sense of direction⁷ or on cues from outside the pool. In any case, the results from the test trial are consistent with the proposal that hippocampally lesioned and normal rats used a heading vector, based on the landmark, to identify the position of the platform.

Two groups of rats that had not been operated on were used in a second experiment to test further our proposals concerning the effects of the hippocampal damage. Group 'session' was trained in the same way as the control group of the previous study. Group 'trial' received similar training, except that the platform and landmark

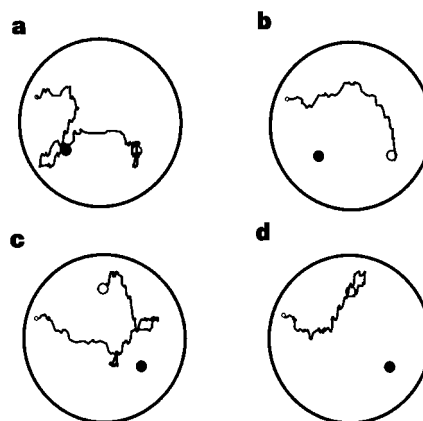


Figure 4 The paths taken by representative normal (a, c) and hippocampally damaged (b, d) rats in the first trials of session 7 (a, b) and session 8 (c, d). The position of the platform in the previous session is indicated by a filled circle, and in the present session by a large open circle; the start point is shown by a small open circle.

were moved, as one, from trial to trial, rather than from session to session. Moving the platform from trial to trial makes it impossible to find by reference to the room cues alone, which will have two consequences. First, group 'trial' cannot show the within-session improvement in performance that would be expected for group 'session'. Second, the tendency to search in the region in which the platform had been located previously at the start of each session should be less for group 'trial' than for group 'session'. The improvement in performance in the first trial of successive sessions should thus be more marked for group 'trial' than for group 'session'. In other words, if our analysis of the performance of the hippocampally damaged rats is correct, moving the platform from trial to trial should have exactly the same effect as a hippocampal lesion.

In support of the above prediction, the results for group 'trial' and group 'session' (Fig. 5) show a striking similarity to those for, respectively, the hippocampally damaged and control groups (Fig. 3). Group 'trial' and the hippocampally damaged rats took less time to find the platform in the first trial of most sessions than did their respective control groups. Furthermore, group 'trial' and the hippocampally damaged rats both failed to show the within-session improvement in performance that was characteristic of the other two groups. An ANOVA using individual mean escape latencies for the 11 sessions combined showed that the latency to escape from the pool was significantly longer in trial 1 than trial 4 for group 'session' ($F = 130.44$, d.f. = 1/14), but not group 'trial' ($F = 3.25$, d.f. = 1/14). There was also a significant difference between the escape latencies for the two groups in trial 1 and trial 4 (F values > 11.24 , d.f. = 1/28).

The gradual improvement in performance in the first trial of each session for normal rats in experiment 1, and for group 'session' in experiment 2, shows that more reliance was placed on the heading-vector strategy, and less on the cognitive-map strategy, as training progressed. According to theories of learning this change would occur because stimuli, or in this case navigational strategies, must compete for the control they acquire over behaviour⁸. At first, animals may be predisposed to adopt a cognitive-map strategy, based on cues outside the pool. However, because the platform can be found more reliably by reference to a landmark-based heading

vector than by a cognitive map, the theories predict that progressively more reliance will be placed on the heading vector. As a result of this switch in strategy, rats should eventually find the platform swiftly at the start of a session, and when the platform and landmark occupy a new position.

Both experiments show that normal rats escaped from the pool by means of two different navigational strategies when the platform remained in the same place for all the trials of a session. One strategy was based entirely on cues that surrounded the pool, and could have involved information about the geometric relationship between the platform and a set of these cues. In other words, rats used a cognitive map to define the position of the platform. The second strategy involved the landmark within the pool, and required rats to learn the direction and distance of the platform from the landmark. That is, rats may also have used a heading vector to define the position of the platform. The results from experiment 1 show that damage to the hippocampus disrupts the first of these strategies, but not the second. □

Methods

The Morris pool was in a well-lit rectangular room with brightly coloured posters on the wall. The lowest point of the spherical black landmark was 2 cm below the surface of the water. For every training trial the midpoint of the 20-cm gap between the sides of the platform and the landmark was located at the midpoint of the radii that pointed towards the eight main points of the compass (Fig. 1). A different release point on the edge of the pool was randomly selected for each training trial. Rats remained on the platform for 30 s once they had climbed onto it. Further details of the method can be found in ref. 4.

Received 29 May; accepted 28 August 1998.

1. O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map* (Clarendon, Oxford, 1978).
2. Collett, T. S., Cartwright, B. A. & Smith, B. A. Landmark learning and visuo-spatial memories in gerbils. *J. Comp. Physiol. A* **158**, 835–851 (1986).
3. Morris, R. G. M., Garrud, P., Rawlins, J. N. P. & O'Keefe, J. Place navigation impaired in rats with hippocampal lesions. *Nature* **297**, 681–683 (1982).
4. Roberts, A. D. L. & Pearce, J. M. Control of spatial behavior by an unstable landmark. *J. Exp. Psychol. Anim. Behav. Proc.* **24**, 172–184 (1998).
5. Jarrard, L. E. On the use of ibotenic acid to lesion selectively different components of the hippocampal formation. *J. Neurosci. Methods* **29**, 251–259 (1989).
6. Morris, R. G. M. Spatial localization does not require the presence of local cue. *Learning Motiv.* **12**, 239–260 (1981).
7. Galistel, C. R. *The Organization of Learning* (MIT Press, Cambridge, MA, 1990).
8. Rescorla, R. A. & Wagner, A. R. in *Classical Conditioning II: Current Research and Theory* (eds Black, A. H. & Prokasy, W. E.) 64–99 (Appleton-Century-Crofts, New York, 1972).

Acknowledgements. We thank R. Honey, E. S. Redhead and R. G. M. Morris for advice and discussions. This work was supported by a BBSRC studentship, and by a grant from the UK Biotechnology and Biological Sciences Research Council.

Correspondence and request for materials should be addressed to J.M.P. (e-mail: pearcejm@cardiff.ac.uk).

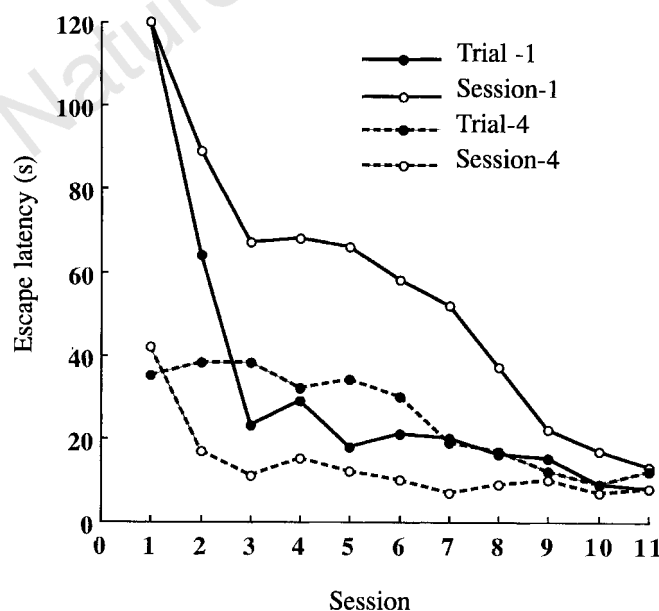


Figure 5 Mean latency of escape in the first and fourth trials of each of the 11 sessions of training, for a group trained with the platform in the same place throughout each session (Session-1 and Session-4) and for a group trained with the platform moved after each trial (Trial-1 and Trial-4).

The anti-inflammatory agents aspirin and salicylate inhibit the activity of I κ B kinase- β

Min-Jean Yin, Yumi Yamamoto & Richard B. Gaynor

Division of Hematology-Oncology, Department of Medicine, Harold Simmons Cancer Center, University of Texas Southwestern Medical Center, Dallas, Texas 75235-8594, USA

NF- κ B comprises a family of cellular transcription factors that are involved in the inducible expression of a variety of cellular genes that regulate the inflammatory response^{1,2}. NF- κ B is sequestered in the cytoplasm by inhibitory proteins, I κ B, which are phosphorylated by a cellular kinase complex known as IKK. IKK is made up of two kinases, IKK- α and IKK- β , which phosphorylate I κ B, leading to its degradation and translocation of NF- κ B to the nucleus^{3–9}. IKK kinase activity is stimulated when cells are

exposed to the cytokine TNF- α or by overexpression of the cellular kinases MEKK1 and NIK^{10,11}. Here we demonstrate that the anti-inflammatory agents aspirin and sodium salicylate specifically inhibit IKK- β activity *in vitro* and *in vivo*. The mechanism of aspirin and sodium salicylate inhibition is due to binding of these agents to IKK- β to reduce ATP binding. Our results indicate that the anti-inflammatory properties of aspirin and salicylate are mediated in part by their specific inhibition of IKK- β , thereby preventing activation by NF- κ B of genes involved in the pathogenesis of the inflammatory response.

Non-steroidal anti-inflammatory drugs (NSAID) such as aspirin, sodium salicylate, and indomethacin exert their anti-inflammatory effects at least in part by inhibition of the enzyme cyclooxygenase^{12,13}. Aspirin and sodium salicylate can also inhibit the NF- κ B pathway which is involved in the pathogenesis of the inflammatory response^{14,15}. We have looked for cellular targets that regulate NF- κ B and could potentially be inhibited by these anti-inflammatory agents. First we assayed the effect of aspirin and sodium salicylate on inhibiting TNF- α and NIK activation of NF- κ B-mediated gene expression^{10,16}. At concentrations (1–5 mM) measured in the serum of patients treated with aspirin and sodium salicylate for chronic inflammatory diseases¹⁷, these agents strongly inhibited NIK- and TNF- α -activated gene expression from the two NF- κ B sites in a human immunodeficiency virus (HIV-1) long terminal repeat (LTR) chloramphenicol acetyltransferase (CAT) reporter plasmid that was trans-

ected into Jurkat cells (Fig. 1a). Aspirin and salicylate also strongly inhibited NF- κ B activation by the human T-lymphotropic virus (HTLV-1) transactivator Tax and the MAP3 kinase MEKK1 (ref. 11) (Fig. 1b). Neither NIK nor TNF- α activated gene expression from an HIV-1 LTR reporter construct containing mutated NF- κ B sites (data not shown). In contrast to the potent inhibition of aspirin and salicylate on NF- κ B activation, neither dexamethasone^{18,19} nor the cyclooxygenase inhibitor indomethacin¹³ prevented NF- κ B activation by TNF- α , NIK, Tax or MEKK1 (Fig. 1a, b), indicating that aspirin and sodium salicylate inhibition of the NF- κ B pathway might not strictly depend on inhibition of prostaglandin synthesis. Finally, we tested whether either aspirin or salicylate could inhibit activation of gene expression by other kinase pathways. The adenovirus early-region-3 promoter, whose gene expression is activated by protein kinase A phosphorylation of CREB, was not inhibited by aspirin and salicylate treatment in the presence of forskolin²⁰ (Fig. 1c). This indicated that aspirin targets only a subset of kinase pathways that activate gene expression.

One of the critical steps in the activation of the NF- κ B pathway is phosphorylation by IKK of I κ B, which leads to its degradation^{5–9}. Kinase activity was assayed with a fusion GST-I κ B α substrate (where GST is glutathione S-transferase) using IKK immunoprecipitated from cells treated with TNF- α in both the presence or absence of anti-inflammatory agents. Aspirin and sodium salicylate, but not dexamethasone or indomethacin, resulted in a 75% inhibition of endogenous IKK kinase activity (Fig. 1d, top). IKK did not phosphorylate a GST-I κ B α substrate mutated at serine residues 32 and 36, which are required for I κ B α phosphorylation and subsequent degradation (data not shown). Although treatment of HeLa cells with aspirin, sodium salicylate or indomethacin decreased prostaglandin levels, only aspirin and sodium salicylate prevented

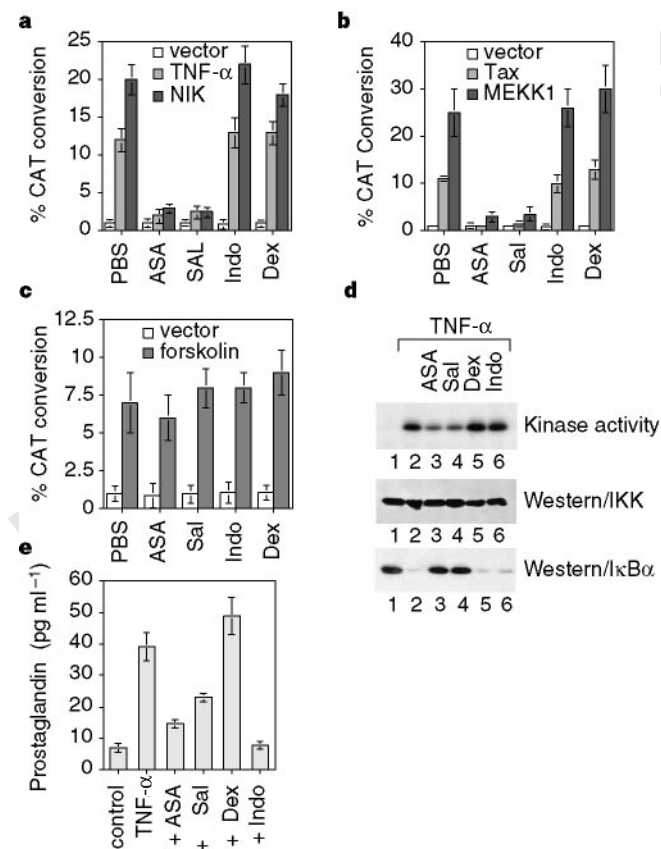


Figure 1 Aspirin inhibits NF- κ B activation and IKK kinase activity. Jurkat cells were transfected with an HIV-1 LTR CAT construct and treated with **a**, TNF- α (20 ng ml⁻¹) or an NIK expression vector; **b**, tax or a MEKK1 expression vector; or **c**, an E3CAT reporter in the presence of forskolin (25 μ M). Cells were treated with either phosphate-buffered saline or agent (ASA, aspirin; SAL, sodium salicylate; Indo, indomethacin; Dex, dexamethasone) and the fold increase in gene expression in the presence of different activators was calculated by phosphorimager quantification. **d**, TNF- α -treated HeLa cell extracts were assayed for endogenous IKK kinase activity (top) and western blot analysis for either IKK- α (middle) or I κ B α (bottom). **e**, Tissue culture medium in **d** was assayed for secreted prostaglandins.

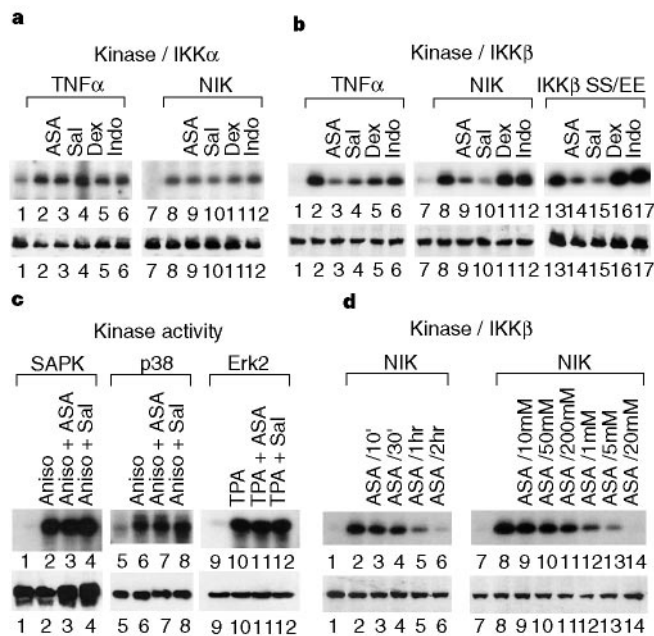


Figure 2 Aspirin specifically inhibits IKK- β kinase activity. Epitope-tagged expression vectors for **a**, IKK- α ; **b**, IKK- β ; **c**, SAPK, p38 and Erk2; or **d**, IKK- β were transfected into COS cells and treated with either TNF- α or NIK, while SAPK and p38 were treated with anisomycin and Erk2 with phorbol esters following 2-h treatment with the indicated anti-inflammatory agents. Immunoprecipitated kinases were assayed with substrates including: **a**, **b**, **d**, GST-I κ B α (aa 1–54); or **c**, GST-Jun (aa 1–169) (lanes 1–4), GST-ATF2 (aa 1–254) (lanes 5–8) or myelin basic protein (lanes 9–12). **d**, Aspirin (5 mM) was added to cells transfected with IKK- β for various times as indicated (lanes 3–6), or different concentrations of aspirin were added for 2 h before collecting cells (lanes 9–14) and IKK- β kinase activity was determined. Western blot analysis of each of these immunoprecipitated kinases is shown in the lower panels.

TNF- α induction of endogenous IKK activity (Fig. 1d, e). Thus, a reduction in prostaglandin synthesis alone was not sufficient to inhibit IKK activity.

To determine whether the activity of either IKK- α or IKK- β could be specifically inhibited by treatment of cells with aspirin or salicylate, expression vectors containing epitope-tagged IKK- α or IKK- β were transfected into COS cells in the presence of either NIK or TNF- α . Both IKK- α and IKK- β activity was increased by treatment of cells with TNF- α and NIK (Fig. 2a, b, lanes 1 and 2). Aspirin and salicylate inhibited IKK- β but not IKK- α kinase activity (Fig. 2a, b, lanes 3 and 4). A constitutively active IKK- β kinase (amino-acid residue mutations S177E and S181E, SS $\rightarrow$ EE)⁷ was also inhibited by aspirin and sodium salicylate treatment (Fig. 2b, lanes 14 and 15), suggesting that these agents may inhibit IKK- β rather than upstream regulatory kinases. Neither aspirin nor sodium salicylate inhibited the activity of the stress-activated protein kinases SAPK, p38 and Erk2 (refs 20–24) (Fig. 2c, lanes 1–12). Finally, we determined the time course and dose dependence *in vivo* of aspirin treatment of cells transfected with an IKK- β expression vector. NIK activated IKK- β kinase activity 40-fold and 50% inhibition of IKK- β kinase activity was seen with aspirin treatment for 30 min (Fig. 2d, lanes 2 and 4) and with an aspirin concentration of 50 μ M (Fig. 2d, lane 10). These results supported the conclusion that aspirin and salicylate selectively inhibit IKK- β activity.

Next we investigated whether aspirin could inhibit IKK- α , IKK- β or IKK- β (SS $\rightarrow$ EE) kinase activity following immunoprecipitation and *in vitro* incubation with either aspirin (Fig. 3a, lanes 1–12) or indomethacin (Fig. 3a, lanes 13–18) for the times indicated. Aspirin treatment did not alter IKK- α activity when produced alone or in the presence of NIK (Fig. 3a, lanes 1–6). In contrast, the kinase activity of both wild-type IKK- β and the constitutively active IKK- β (SS $\rightarrow$ EE) were both decreased by more than 90% when treated *in vitro* with aspirin (Fig. 3a, lanes 7–12). *In vitro* treatment

with indomethacin did not decrease the kinase activity of either the wild-type or the constitutively active IKK- β (SS $\rightarrow$ EE) (Fig. 3a, lanes 13–18). Consistent with our *in vivo* results, *in vitro* treatment with aspirin did not alter the kinase activity of immunoprecipitated SAPK, p38 or Erk2 (Fig. 3b, lanes 1–12). Aspirin also inhibited the *in vitro* kinase activity of baculovirus-expressed IKK- β (Fig. 3c, lanes 3–6) but not IKK- α (Fig. 3c, lanes 1–3). The half-maximal inhibitory concentration (IC₅₀) of aspirin required to inhibit prostaglandin synthesis has been reported to range from 50–100 μ M^{25,26}. The IC₅₀ for *in vivo* aspirin treatment required to inhibit endogenous IKK kinase activity was 80 μ M (Fig. 4a). Aspirin

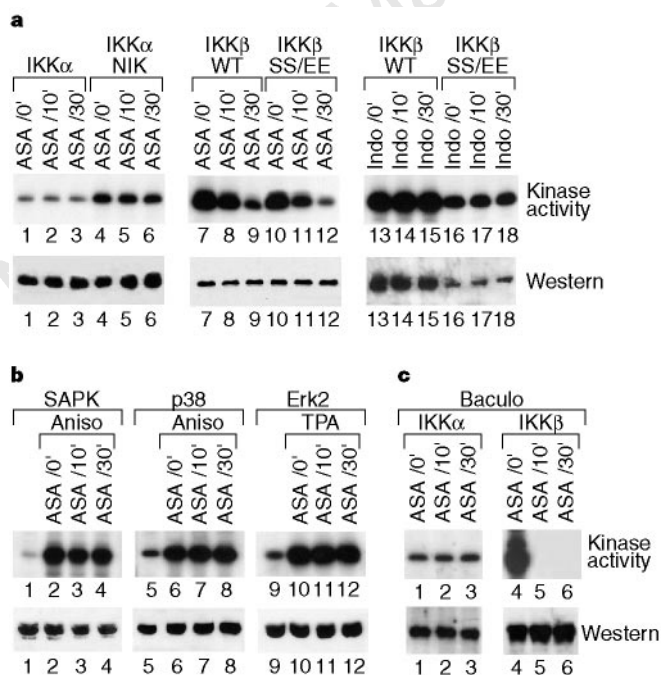


Figure 3 Aspirin inhibits IKK- β kinase activity *in vitro*. COS cells were transfected with: **a**, epitope-tagged IKK- α or IKK- β , or **b**, SAPK, p38, or Erk2 cDNAs treated with either anisomycin or phorbol esters. Kinases were immunoprecipitated and incubated *in vitro* for 0, 10 or 30 min with either 1 mM aspirin or 25 μ M indomethacin, followed by kinase assays. **c**, Baculovirus-produced IKK- α (lanes 1–3) or IKK- β (lanes 4–6) proteins were also incubated *in vitro* with 1 mM aspirin for either 10 or 30 min and assayed for kinase activity. Lower panels, western blot analysis of each immunoprecipitate. WT, wild type; SE/EE, mutants.

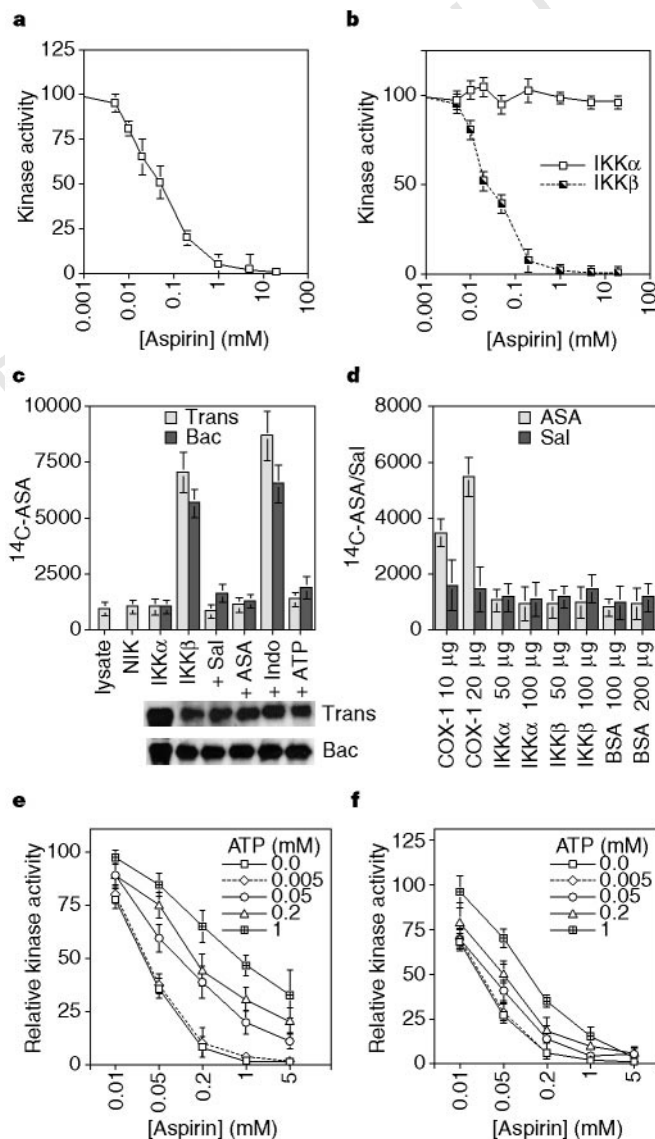


Figure 4 Aspirin and salicylate compete with ATP for binding to IKK- β . **a**, HeLa cells were treated with various concentrations of aspirin *in vivo* for 2 h before TNF- α treatment and assay of endogenous IKK activity. **b**, Baculovirus-produced IKK- α and IKK- β proteins were incubated *in vitro* with various concentrations of aspirin and assayed for kinase activity. **c**, NIK, IKK- α or IKK- β proteins produced following either transfection (Trans) or baculovirus (Bac) expression were immunoprecipitated and incubated with ¹⁴C-aspirin in the presence of a 500-fold molar excess of unlabelled competitors and binding was quantified; western-blot analysis of the immunoprecipitated IKK proteins is shown in the lower panel. **d**, Purified IKK- α , IKK- β or COX-1 proteins were incubated with either ¹⁴C-aspirin or ¹⁴C-salicylic acid, precipitated with 20% TCA and quantified by β counting. **e**, **f**, Purified IKK- β was incubated with **e**, different concentrations of aspirin and ATP added together, or **f**, preincubated with aspirin for 30 min before addition of ATP and assayed for kinase activity.

also inhibited the *in vitro* kinase activity of baculovirus-expressed IKK- β but not IKK- α at 30–50 μ M (Fig. 4b). Thus, the IC₅₀ for *in vitro* aspirin inhibition of IKK- β correlated with the IC₅₀ for *in vivo* aspirin inhibition of endogenous IKK.

We next assayed the ability of aspirin and salicylate to bind to IKK- α or IKK- β isolated from either transfected cells or purified following baculovirus expression. Sevenfold more ¹⁴C-aspirin bound to IKK- β than to either IKK- α or NIK (Fig. 4c). Moreover, the binding of ¹⁴C-aspirin or [7-¹⁴C]salicylic acid (data not shown) to IKK- β was competed by a 500-fold molar excess of either unlabelled salicylate or aspirin, but not indomethacin (Fig. 4c). A 500-fold molar excess of ATP was also able to compete with the binding of ¹⁴C-aspirin to IKK- β , indicating that ATP and aspirin competed for binding to IKK- β (Fig. 4c). To determine whether aspirin or salicylate was covalently bound to IKK- β , we used trichloroacetic acid (TCA) to precipitate recombinant IKK- α and IKK- β bound to either ¹⁴C-salicylate or ¹⁴C-aspirin (Fig. 4d). TCA precipitation did not reveal significant amounts of ¹⁴C-aspirin or ¹⁴C-salicylate remaining bound to precipitated IKK- β . In contrast, the cyclooxygenase COX-1, which binds covalently to aspirin but not salicylate, bound to ¹⁴C-aspirin but not ¹⁴C-salicylate (Fig. 4d)^{25,26}. These results indicate that aspirin and salicylate do not bind covalently to IKK- β .

To confirm that aspirin and salicylate compete with ATP for binding to IKK- β , we preincubated increasing concentrations of ATP and aspirin with purified IKK- β before assaying for kinase activity. The IC₅₀ of aspirin required to inhibit IKK- β shifted with increasing ATP concentration, indicating that aspirin was a competitive inhibitor of ATP binding to IKK- β (Fig. 4e). When aspirin was preincubated with IKK- β for 30 min before addition of ATP, the IC₅₀ of aspirin did not markedly shift with increasing ATP concentration (Fig. 4f), indicating that the binding of aspirin to IKK- β was either irreversible or very slowly reversible, although this binding is not covalent. As NF- κ B is a key cellular regulator of the inflammatory response, inhibition of the nuclear translocation of NF- κ B should suppress the host inflammatory response. Our identification of IKK- β as a target for sodium salicylate and aspirin should help in the design of aspirin-like agents to inhibit the inflammatory response even more effectively. □

Methods

Cell culture and transfections. COS and HeLa cells were transfected with Eugene 6 (Boehringer–Mannheim); Jurkat cells were transfected with DEAE-dextran. Cells were collected 24 h post-transfection in the absence or presence of aspirin (5 mM), sodium salicylate (5 mM), dexamethasone (10 μ M) or indomethacin (25 μ M). The HIV1-LTR-CAT and E3-CAT reporter constructs^{11,20}, and the epitope-tagged IKK- α (HA), IKK- β (Flag), NIK (c-Myc), Tax, MEKK1, p38 (HA), SAPK (Myc), Erk2 (Myc) have been described^{11,21–24}. TNF- α (20 ng ml⁻¹) was added to cells for 10 min before collection to stimulate IKK kinase activity and for 20 h post-transfection for assay of NF- κ B-mediated gene expression. Cells transfected with SAPK and p38 cDNAs were treated with anisomycin (10 μ g ml⁻¹) for 30 min before collection; cells transfected with the Erk2 cDNA were pretreated with TPA (12-O-tetradecanoylphorbol-13-acetate; 50 ng ml⁻¹) for 30 min²⁴ to activate these kinases. Both aspirin (acetyl salicylic acid) and sodium salicylate (Sigma) were dissolved in 0.05 M Tris-HCl to prepare 1.0 M stock solutions; dexamethasone and forskolin (Sigma) were prepared as described¹⁸. Supernatants from cells were applied to a C18 minicolumn and assayed for prostaglandin using an ELISA kit (Amersham).

Kinase assays. Lysates (200 μ g protein) were prepared from transfected cells and incubated with antibody (anti-Flag (M2), anti-HA (12CA5), or anti-Myc) at 4 °C for 1 h and 20 μ l protein A-agarose was added for 1 h. After extensively washing the immunoprecipitates, kinase assays were done as described¹¹. For *in vitro* kinase assay, aspirin was added into the washed immunoprecipitates for 30 min at 4 °C before the kinase reaction. Mixtures were subjected to SDS-PAGE and autoradiography and quantified by phosphorimager analysis.

Calculation of the IC₅₀ of aspirin. To assay endogenous IKK activity, cell lysates (200 μ g protein) were immunoprecipitated with a rabbit polyclonal antibody directed against IKK- α (Santa Cruz) that immunoprecipitates the

IKK- α /IKK- β heterodimer, followed by assay of kinase activity. Polyhistidine and Flag-tagged IKK- α and IKK- β proteins were produced by baculovirus expression and purified using nickel-agarose chromatography. Purified proteins (500 μ g) were immunoprecipitated using 12CA5 monoclonal antibody, divided into ten equal fractions and each was treated with a different concentration of aspirin or sodium salicylate for 30 min on ice. Kinase activity was then assayed and quantified by phosphorimager; aspirin inhibition of kinase activity was calculated and plotted against aspirin concentration.

IKK binding to ¹⁴C-salicylate and ¹⁴C-aspirin binding. Proteins (200 μ g) isolated from baculovirus-expressed and purified IKK- α and IKK- β proteins or cells transfected with IKK- α or IKK- β cDNAs were immunoprecipitated with epitope-specific monoclonal antibodies and then incubated with 500 μ l binding buffer containing 100 mM NaCl, 50 mM Tris, pH 7.5, 10 mg ml⁻¹ BSA, protease inhibitors, and 2 μ Ci of either acetyl salicylic ¹⁴C-carboxylic acid or [7-¹⁴C]salicylic acid (40–60 mCi mmol⁻¹). A 500-fold molar excess (36 mM) of aspirin, sodium salicylate, indomethacin or ATP was added to each immunoprecipitate and incubated at 4 °C for 30 min. Immunoprecipitates were then washed extensively with binding buffer and the amount of ¹⁴C-salicylate or ¹⁴C-aspirin bound was quantified by β -counting. Immunoprecipitates were also incubated with 20% TCA, precipitates were isolated by centrifugation and dissolved in 1 M NaOH. The amount of ¹⁴C-aspirin and ¹⁴C-salicylate in the protein precipitates was quantified by β counting. Either 10 or 20 μ g COX-1 protein (from Cayman Chemical) was used in binding reactions with IKK proteins.

Received 13 August; accepted 17 September 1998.

- Baeuerle, P. A. & Baltimore, D. NF- κ B: Ten years after. *Cell* **87**, 13–20 (1996).
- Baldwin, J. A. S. The NF- κ B and I κ B proteins: new discoveries and insights. *Annu. Rev. Immunol.* **14**, 649–681 (1996).
- Brown, K., Gerstberger, S., Carlson, L., Franzoso, G. & Siebenlist, U. Control of I κ B α proteolysis by site specific signal induced phosphorylation. *Science* **267**, 1485–1488 (1995).
- Chen, Z. *et al.* Signal-induced site-specific phosphorylation targets I κ B α to the ubiquitin-proteasome pathway. *Genes Dev.* **9**, 1586–1597 (1995).
- DiDonato, J. A., Hayakawa, M., Rothwarf, D. M., Zandi, E. & Karin, M. A cytokine-responsive I κ B kinase that activates the transcription factor NF- κ B. *Nature* **388**, 548–554 (1997).
- Régnier, C. H. *et al.* Identification and characterization of an I κ B kinase. *Cell* **90**, 373–383 (1997).
- Mercurio, F. *et al.* IKK-1 and IKK-2: cytokine-activated I κ B kinases essential for NF- κ B activation. *Science* **278**, 860–866 (1997).
- Zandi, E., Rothwarf, D. M., Delhase, M., Hayakawa, M. & Karin, M. The I κ B kinase complex (IKK) contains two kinase subunits, IKK α and IKK β , necessary for I κ B phosphorylation and NF- κ B activation. *Cell* **91**, 243–252 (1997).
- Woronicz, J. D., Gao, X., Cao, Z., Rothe, M. & Goeddel, D. V. I κ B kinase- β : NF- κ B activation and complex formation with I κ B kinase- α and NIK. *Science* **278**, 866–869 (1997).
- Malinin, N. L., Boldin, M. P., Kovalenko, A. V. & Wallach, D. MAP3K-related kinase involved in NF- κ B induction by TNF, CD95 and IL-1. *Nature* **385**, 540–544 (1997).
- Yin, M.-J. *et al.* HTLV-1 Tax protein binds to MEKK1 to stimulate I κ B kinase activity and NF- κ B activation. *Cell* **93**, 875–884 (1998).
- Vane, J. R. Inhibition of prostaglandin synthesis as a mechanism of action of the aspirin-like drugs. *Nature New Biol.* **231**, 232–235 (1971).
- Vane, J. Towards a better aspirin. *Nature* **367**, 215–216 (1994).
- Kopp, E. & Ghosh, S. Inhibition of NF- κ B by sodium salicylate and aspirin. *Science* **265**, 956–959 (1994).
- Grilli, M., Pizzi, M., Memo, M. & Spano, P. Neuroprotection by aspirin and sodium salicylate through blockade of NF- κ B activation. *Science* **274**, 1383–1385 (1996).
- Rothe, M., Sarma, V., Dixit, V. M. & Goeddel, D. V. TRAF2-mediated activation of NF- κ B by TNF receptor 2 and CD40. *Science* **269**, 1424–1427 (1995).
- Insel, P. *The Pharmacological Basis of Therapeutics* (eds Harman, J. G., Molinoff, P. B., Ruddon, R. W. & Gilman, A. G.) (McGraw-Hill, New York, 1996).
- Auphan, N., DiDonato, J. A., Rosette, C., Helmberg, A. & Karin, M. Immunosuppression by glucocorticoids: Inhibition of NF- κ B activity through induction of I κ B synthesis. *Science* **270**, 286–290 (1995).
- Scheinman, R. I., Cogswell, P. C., Lofquist, A. K. & Baldwin, A. S. Role of transcriptional activation of I κ B α in mediation of immunosuppression by glucocorticoids. *Science* **270**, 283–286 (1995).
- Tan, T. H., Jia, R. & Roeder, R. G. Utilization of signal transduction pathway by the human T-cell leukemia virus type I transcriptional activator tax. *J. Virol.* **63**, 3761–3768 (1989).
- Boulton, T. G. *et al.* ERKs: a family of protein-serine/threonine kinases that are activated and tyrosine phosphorylated in response to insulin and NGF. *Cell* **65**, 663–675 (1991).
- Khokhlatchev, A. *et al.* Reconstitution of mitogen-activated protein kinase phosphorylation cascades in bacteria. Efficient synthesis of active protein kinases. *J. Biol. Chem.* **272**, 11057–11062 (1997).
- Xu, S. & Cobb, M. H. MEKK1 binds directly to the c-Jun N-terminal kinases/stress-activated protein kinases. *J. Biol. Chem.* **272**, 32056–32060 (1997).
- Kyriakis, J. M. *et al.* The stress-activated protein kinase subfamily of c-Jun kinases. *Nature* **369**, 156–160 (1994).
- Mitchell, J. A., Akaraseenont, P., Thiermermann, C., Flower, R. J. & Vane, J. R. Selectivity of nonsteroidal antiinflammatory drugs as inhibitors of constitutive and inducible cyclooxygenase. *Proc. Natl Acad. Sci. USA* **90**, 11693–11697 (1994).
- Kalgtutkar, A. S. *et al.* Aspirin-like molecule that covalently inactivates cyclooxygenase-2. *Science* **280**, 1268–1270 (1998).

Acknowledgements. We thank F. Mercurio for IKK- α and IKK- β cDNAs, M. Cobb and E. Ross for discussion, and S. Johnson and E. Pasternack respectively for preparing the manuscript and figures. This work was supported by grants from the NIH, the Council for Tobacco Research and the Robert Welch Foundation.

Correspondence and requests for materials should be addressed to R.B.G. (e-mail: gaynor@utsw.swmed.edu).

Type III InsP₃ receptor channel stays open in the presence of increased calcium

Robert E. Hagar*, Angela D. Burgstahler†, Michael H. Nathanson† & Barbara E. Ehrlich‡

* Department of Physiology, University of Connecticut Health Center, Farmington, Connecticut 06030, USA

Departments of † Medicine and Cell Biology, and ‡ Pharmacology and Cellular and Molecular Physiology, Yale University, New Haven, Connecticut 06520, USA

The inositol 1,4,5-trisphosphate receptor (InsP₃R) is the main calcium(Ca²⁺) release channel in most tissues. Three isoforms have been identified^{1–6}, but only types I and II InsP₃R have been characterized^{7,8}. Here we examine the functional properties of the type III InsP₃R because this receptor is restricted to the trigger zone from which Ca²⁺ waves originate^{9–11} and it has distinctive InsP₃-binding properties^{12,13}. We find that type III InsP₃R forms Ca²⁺ channels with single-channel currents that are similar to those of type I InsP₃R; however, the open probability of type III InsP₃R isoform increases monotonically with increased cytoplasmic Ca²⁺ concentration, whereas the type I isoform has a bell-shaped dependence on cytoplasmic Ca²⁺. The properties of type III InsP₃R provide positive feedback as Ca²⁺ is released; the lack of negative feedback allows complete Ca²⁺ release from intracellular stores. Thus, activation of type III InsP₃R in cells that express only this isoform results in a single transient, but global, increase in the concentration of cytosolic Ca²⁺. The bell-shaped Ca²⁺-dependence curve of type I InsP₃R is ideal for supporting Ca²⁺ oscillations, whereas the properties of type III InsP₃R are better suited to signal initiation.

When homogenates from RIN-5F cells, rat hepatocytes, and canine cerebellum were probed by western blot analysis with isoform-specific antibodies, RIN-5F cells contained no detectable type I InsP₃R (Fig. 1a, lane 4) but did express type III InsP₃R (Fig. 1b, lane 4). In contrast, both hepatocytes and cerebellum expressed type I InsP₃R (Fig. 1a, lanes 1 and 3), but no detectable type III InsP₃R (Fig. 1b, lanes 1 and 3). Thus, the InsP₃R of RIN-5F cells is almost entirely the type III isoform, consistent with previous reports¹⁴. Immunocytochemistry demonstrated that type III InsP₃R was diffusely distributed (Fig. 1c) and confirmed that type

I InsP₃R expression was minimal (Fig. 1d, e). These findings do not support a previous prediction that type III InsP₃R is preferentially localized near the plasma membrane¹⁵.

To test whether type III InsP₃R forms a Ca²⁺ channel, we incorporated endoplasmic reticulum vesicles from cultured RIN-5F cells into planar lipid bilayers. Single-channel currents were observed through type III InsP₃R. Like type I InsP₃R, type III InsP₃R required InsP₃ to open (Fig. 2a: compare top three and bottom three traces). The single-channel current of type III InsP₃R measured at 0 mV with 55 mM Ba²⁺ on the luminal side of the channel was ~2.5 pA, which is similar to that of type I InsP₃R (ref. 7). The conductances of type III InsP₃R (88 ± 4 pS; Fig. 2b) and type I InsP₃R (85 ± 3 pS)⁷ were also similar.

A central feature of type I InsP₃R is that Ca²⁺ acts as an allosteric regulator when InsP₃ is present. Maximum channel activity occurs when the concentration of free Ca²⁺ is 250 nM; there is a sharp decline in channel activity on either side of the maximum, with complete inhibition when cytosolic Ca²⁺ exceeds 5 μM (Fig. 3b, circles)^{16,17}. This bell-shaped regulation provides amplification of the initial InsP₃ signal, as well as negative-feedback inhibition of further InsP₃-stimulated Ca²⁺ release. The feedback inhibition is particularly important because it provides autoregulation and is essential for Ca²⁺ oscillations and for the propagation of regenerative intracellular Ca²⁺ waves¹⁸.

Does Ca²⁺ regulate type III InsP₃R in the same way? We hypothesized that type III InsP₃R should remain open at high cytoplasmic Ca²⁺ concentration ([Ca²⁺]) because InsP₃ binding to type III InsP₃R is not inhibited by elevated Ca²⁺ (refs 12, 13). To test this idea, we altered the cytoplasmic [Ca²⁺] in the presence of a fixed InsP₃ concentration (2 μM). Like type I InsP₃R, type III InsP₃R was progressively activated as the cytoplasmic [Ca²⁺] increased to 250 nM. However, channel activity for type III InsP₃R remained high even when cytoplasmic [Ca²⁺] was raised to 100 μM (Fig. 3a, bottom traces; Fig. 3b, triangles). The mean open time for type III InsP₃R (6.5 ± 0.62 ms) was nearly identical to the reported value for type I InsP₃R (6.4 ± 1.0 ms)⁷ and was independent of the cytoplasmic Ca²⁺ concentration. Thus, cytoplasmic free Ca²⁺ regulates the

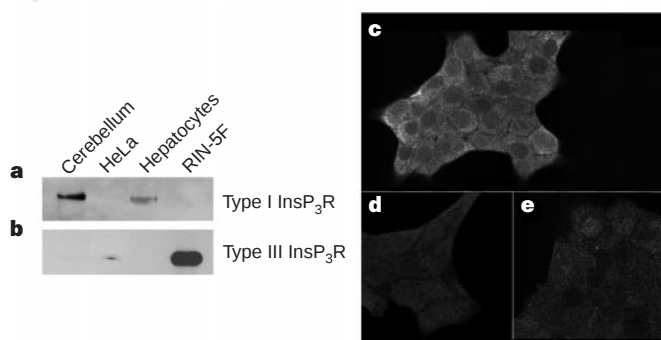


Figure 1 RIN-5F cells preferentially express Type III InsP₃R. **a, b**, Western blots were probed for types I and III InsP₃R (**a** and **b**, respectively). Dog cerebellum and rat hepatocytes were positive controls for type I InsP₃R; HeLa cells were a positive control for type III InsP₃R. Lanes were loaded with 30 μg dog cerebellar microsomes (lane 1), 10 μg HeLa cell lysate (lane 2), 30 μg hepatic microsomes (lane 3), and either 30 μg (**a**) or 5 μg (**b**) of RIN-5F microsomes (lane 4). **c, e**, Cellular distribution of type III (**c**) and type I (**e**) InsP₃R in RIN-5F cells. **d**, Nonspecific binding.

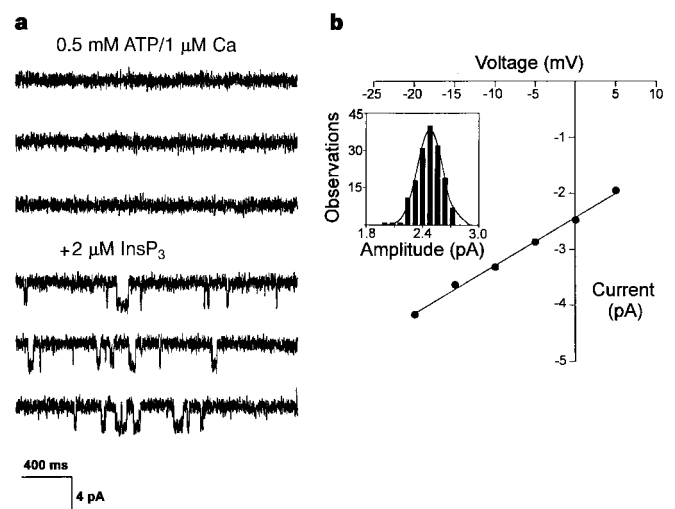


Figure 2 Type III InsP₃R is an InsP₃-gated Ca²⁺ channel. **a**, InsP₃-gated Ca²⁺ channels from endoplasmic reticulum of RIN-5F cells in planar lipid bilayers. In the absence of InsP₃, channel activity was not observed (top three traces). Addition of 2 μM InsP₃ to the cytoplasmic side induced channel activity (bottom three traces). Channel openings are shown as downward deflections from baseline. Ruthenium red (2 μM) was present to block ryanodine receptors. **b**, Current-voltage relationship of type III InsP₃R. Inset shows an amplitude histogram at 0 mV for one experiment. Values plotted in the I-V curve represent the mean for three experiments. Standard errors for data points, which range from 0.02 to 0.05 pA, are too small to be seen.

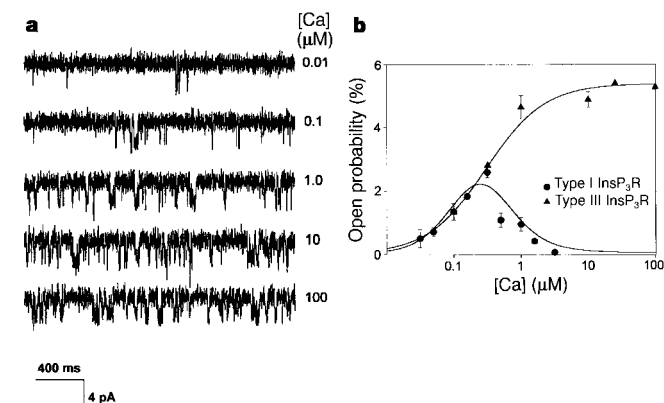


Figure 3 Single channel open probability for type I and type III InsP_3R as a function of Ca^{2+} concentration. **a**, Channel activity for type III InsP_3R in the presence of $2\ \mu\text{M}$ InsP_3 , $0.5\ \text{mM}$ ATP, $0.5\ \text{mM}$ EGTA, $2\ \mu\text{M}$ ruthenium red, and at $0\ \text{mV}$. Channel openings are shown as downward deflections. **b**, Single channel open probability of type I InsP_3R (circles) and type III InsP_3R (triangles). Data points for type I InsP_3R were taken from ref. 17. Data for three experiments are shown for type III InsP_3R . Individual points with error bars are the mean $\pm$ s.e.m. for $n = 2$ (1 and $10\ \mu\text{M}$ Ca^{2+}) or $n = 3$ (0.01 and $0.1\ \mu\text{M}$ Ca^{2+}).

two isoforms differently; although both isoforms show similar Ca^{2+} -dependent activation, Ca^{2+} -dependent inhibition is lacking for type III InsP_3R .

To investigate the physiological importance of this pattern of activation in an intact cell, we monitored the cytoplasmic $[\text{Ca}^{2+}]$ in RIN-5F cells which we stimulated with extracellular ATP that binds to P_{2Y} receptors and increases $[\text{Ca}^{2+}]$ through the InsP_3 cascade^{19,20}. RIN-5F cells responded to ATP stimulation with a single, transient increase in $[\text{Ca}^{2+}]$ (Fig. 4a). In single cells, stimulation with $100\ \mu\text{M}$ ATP induced a $209 \pm 24\%$ increase in Fluo-3 fluorescence relative to baseline; this increase lasted for $15.2 \pm 0.8\ \text{s}$ (Table 1). The response was similar in Ca^{2+} -free medium (Table 1). Similar, but slightly smaller, increases in Fluo-3 fluorescence were seen in cells stimulated with $10\ \mu\text{M}$ or $1\ \mu\text{M}$ ATP. There was no increase in $[\text{Ca}^{2+}]$ in response to stimulation with $0.1\ \mu\text{M}$ ATP (Fig. 4b), and neither sustained nor repetitive increases in $[\text{Ca}^{2+}]$ (that is, Ca^{2+} oscillations) were seen in any of the cells. In contrast, hepatocytes, which contain only types I and II InsP_3R , produced Ca^{2+} oscillations after stimulation with $1\ \mu\text{M}$ ATP (Fig. 4c), as shown previously²¹. RIN-5F cells that were serially stimulated with $10\ \mu\text{M}$ and then $100\ \mu\text{M}$ ATP responded only to the initial exposure to $10\ \mu\text{M}$ ATP ($n = 8$). In addition, thapsigargin ($2\ \mu\text{M}$) had a minimal effect on intracellular $[\text{Ca}^{2+}]$ when RIN-5F cells were pretreated with ATP (Fig. 4d). This finding provides direct evidence that activation of type III InsP_3R drains internal Ca^{2+} stores in RIN-5F cells. Furthermore, the magnitude of ATP-induced Ca^{2+} spikes is significantly greater in RIN-5F cells than in hepatocytes (Fig. 4d, f). Thus, positive feedback of Ca^{2+} on type III InsP_3R in RIN-5F cells

Table 1 Ca^{2+} transients in RIN-5F cells after stimulation by ATP

ATP (μM)	Magnitude (% increase)	Duration (s)
100	209 ± 24 ($n = 23$)	15.2 ± 0.8 ($n = 31$)
100 (Ca^{2+} -free)	200 ± 13 ($n = 25$)	14.8 ± 1.2 ($n = 25$)
10	160 ± 14 ($n = 25$)	16.8 ± 2.1 ($n = 19$)
1	121 ± 15 ($n = 15$)	13.0 ± 2.1 ($n = 8$)
0.1	0 ($n = 32$)	0 ($n = 32$)

RIN-5F cells were stimulated with ATP and changes in cytoplasmic Ca^{2+} in individual cells were measured by Fluo-3 fluorescence using confocal line scanning microscopy. The magnitude (per cent increase over baseline) and duration of the Ca^{2+} transients are expressed as mean $\pm$ s.e.m. (n).

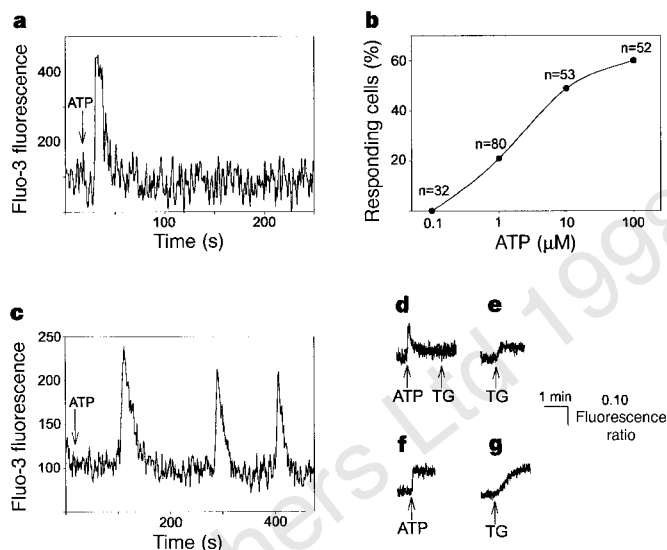


Figure 4 Ca^{2+} signalling patterns in RIN-5F cells and rat hepatocytes. **a**, Stimulation of a RIN-5F cell with $100\ \mu\text{M}$ ATP induces a single intracellular Ca^{2+} transient. **b**, Dose-response curve for RIN-5F cells stimulated with ATP. Values indicate total numbers of cells stimulated. **c**, Stimulation of an hepatocyte with $1\ \mu\text{M}$ ATP induces Ca^{2+} oscillations. **d**, ATP ($100\ \mu\text{M}$) induces a single transient increase in Ca^{2+} in RIN-5F populations (peak change in fluorescence ratio, $\Delta R = 0.20 \pm 0.01$; $n = 6$). Subsequent treatment with thapsigargin (TG; $2\ \mu\text{M}$) has little effect on Ca^{2+} ($\Delta R = 0.03 \pm 0.01$; $n = 6$). **e**, Thapsigargin alone ($2\ \mu\text{M}$) increases Ca^{2+} in RIN-5F cells ($\Delta R = 0.08 \pm 0.01$; $n = 6$). **f**, ATP ($100\ \mu\text{M}$) induces a rapid, sustained increase in Ca^{2+} in hepatocyte populations ($\Delta R = 0.13 \pm 0.02$; $n = 7$). **g**, Thapsigargin ($2\ \mu\text{M}$) increases Ca^{2+} in hepatocytes like ATP alone ($\Delta R = 0.10 \pm 0.02$; $n = 7$).

causes rapid, massive, near-complete Ca^{2+} release, and results in more intense Ca^{2+} spikes that are of shorter duration than those that occur through the type I InsP_3R in hepatocytes.

We next compared subcellular Ca^{2+} release from type III InsP_3R in RIN-5F cells with subcellular Ca^{2+} release from type I InsP_3R in SKHepl cells (a hepatoma cell line that expresses type I but not type III InsP_3R ; our unpublished observation). Small amounts of InsP_3 were released in both cell types by flash photolysis of microinjected caged InsP_3 . Release of InsP_3 in the RIN-5F cells always resulted in all-or-none global Ca^{2+} signalling, although increases in $[\text{Ca}^{2+}]$ began in focal subcellular regions before spreading to encompass the entire cell (Fig. 5a–d). In contrast, non-propagating increases in $[\text{Ca}^{2+}]$ could sometimes be elicited in SKHepl cells by photorelease of minimal amounts of InsP_3 (Fig. 5e–g, i), similar to responses previously seen in pancreatic acinar and HeLa cells and *Xenopus* oocytes^{10,11,22,23}. Photorelease of larger amounts of InsP_3 induced a global Ca^{2+} response in SKHepl cells (Fig. 5h, j). These findings suggest that subcellular Ca^{2+} release via type III InsP_3R results in a positive-feedback cycle that leads to all-or-none Ca^{2+} signalling that spreads throughout the cell, whereas release through type I InsP_3R with low concentrations of InsP_3 results in localized, non-propagating increases in $[\text{Ca}^{2+}]$.

The localized, subcellular Ca^{2+} signals in SKHepl cells lasted from one to several seconds, a result that is typical for the duration of subcellular Ca^{2+} signals described in other mammalian cells^{10,11,24}. Localized increases in $[\text{Ca}^{2+}]$ of shorter duration are routinely seen in *Xenopus* oocytes, in which Ca^{2+} puffs typically last for 300–1,000 ms (ref. 23). Although individual Ca^{2+} puffs are restricted to a small fraction of the total volume of an oocyte²³, they occupy a region that is larger than an entire mammalian cell. We believe that the localized elevation in $[\text{Ca}^{2+}]$ in SKHepl cells is a reasonable mammalian cell equivalent of the localized Ca^{2+} signals described in oocytes.

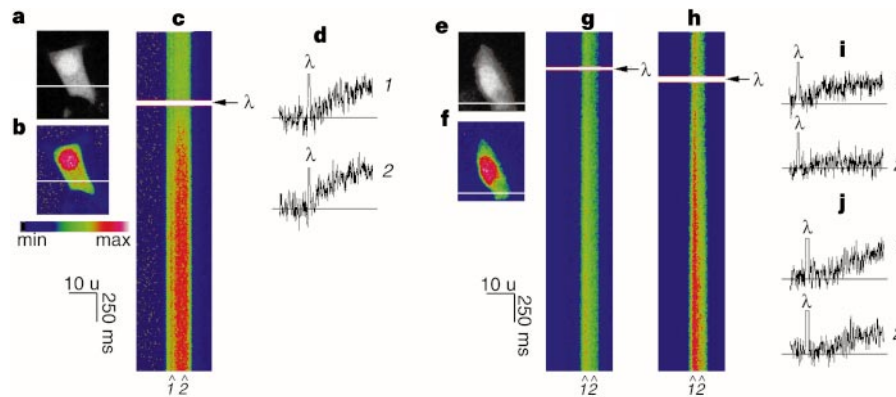


Figure 5 Subcellular Ca^{2+} release differs between RIN-5F and SKHep1 cells. **a**, Confocal image of a RIN-5F cell. **b**, Pseudocolour image of the cell loaded with Fluo-3. The same pseudocolour scale was used for **b**, **c**, and **f-h**. **c**, Line scan collected along the line indicated in **b** after flash photolysis (λ) of caged InsP_3 . Ca^{2+} increases throughout the cell after photorelease of InsP_3 (representative of 13 experiments with flash duration 50–100 ms). No Ca^{2+} increase was detected in 6 experiments with flash duration <50 ms. Spatial resolution, $0.26 \mu\text{m}$ per pixel; temporal resolution, 6 ms per pixel. **d**, Release of caged InsP_3 results in a global increase in Ca^{2+} . Trace duration in **d**, **i** and **j** is 2 s. **e**, **f**, Confocal (**e**) and

pseudocolour (**f**) images of an SKHep1 cell. **g**, **h**, Confocal line scans of the cell during 30 and 60 ms flashes to photolyse caged InsP_3 . A response is detected in only the left side of the cell after a short flash (**g**), and throughout the cell after a long flash (**h**). Responses were absent or minimal in only part of an SKHep1 cell in 7 experiments (flash duration, 3–50 ms), whereas global Ca^{2+} increases were seen in 12 experiments (flash duration, 5–60 ms). **i**, An increase occurs on the left side of the cell in **g** (1), but not on the right (2). **j**, After more photorelease of InsP_3 (**h**), similar Ca^{2+} increases occur at the same two subcellular locations.

Ca^{2+} concentrations in RIN-5F cells returned nearly to baseline after stimulation with ATP (Fig. 4a, d), although depletion of intracellular Ca^{2+} stores in other cells activates a robust store-operated Ca^{2+} current (I_{soc})²⁵. Our findings suggest that I_{soc} is either small or absent in RIN-5F cells. We therefore stimulated RIN-5F cells with thapsigargin ($2 \mu\text{M}$) to deplete Ca^{2+} stores, and then removed Ca^{2+} from the medium. Thapsigargin induced a small but sustained increase in intracellular $[\text{Ca}^{2+}]$ (Fig. 4e); this increase was abolished by subsequent addition of extracellular EGTA. When RIN-5F cells were stimulated with thapsigargin in the absence of extracellular Ca^{2+} , there was no sustained increase in intracellular $[\text{Ca}^{2+}]$. Addition of thapsigargin to hepatocytes, in which I_{soc} has been described²⁵, induced a larger increase in intracellular $[\text{Ca}^{2+}]$ (Fig. 4g); the level of Ca^{2+} also returned to baseline in response to extracellular EGTA. These findings indicate that RIN-5F cells contain less I_{soc} than hepatocytes, which is interesting in light of a suggestion that the primary role of type III InsP_3R is to initiate I_{soc} (ref. 15) rather than Ca^{2+} transients. This hypothesis was based upon heterologous expression studies in *Xenopus* oocytes in which the protein seems to be targeted to the cell surface and may differ from the distribution found in RIN-5F cells (Fig. 1b) and other mammalian cells^{5,9}.

The high degree of homology among InsP_3R isoforms indicates that many of their functional properties ought to be comparable. We found that properties such as activation by InsP_3 , the magnitude of the single-channel current, and activation by concentrations of Ca^{2+} less than 250 nM were quite similar. In addition, the sustained activity of type III InsP_3R at raised Ca^{2+} concentrations is similar to that reported for type I InsP_3R in the presence of high concentrations of InsP_3 ($180 \mu\text{M}$)¹⁷. In both cases, there is a lack of Ca^{2+} -dependent inhibition, and the channel remains open even when the cytoplasmic Ca^{2+} exceeds $50 \mu\text{M}$. In the presence of low concentrations of InsP_3 ($<5 \mu\text{M}$), however, the two isoforms have fundamentally distinct responses to cytoplasmic $[\text{Ca}^{2+}]$ greater than 250 nM (Fig. 3). If type III InsP_3R does not show Ca^{2+} -dependent inhibition, what closes the channel? The primary level of regulation may be the control of InsP_3 generation and degradation because type III, like type I, InsP_3R only opens when InsP_3 is present. Further regulation of intracellular Ca^{2+} release by type III InsP_3R can occur by locally depleting the intracellular Ca^{2+} stores. Other mechanisms shown to be important for type I InsP_3R , such as phosphorylation

and regulation by associated proteins²⁶, may also modulate the activity of type III InsP_3R .

Multiple isoforms of the InsP_3R are expressed in a variety of cell types¹⁴, but the physiological significance of this was unclear. The functional differences between type I and type III isoforms of the InsP_3R now indicate that each has a special role in the cell. Type I InsP_3R , with both Ca^{2+} -dependent activation and inhibition, is well suited for establishing Ca^{2+} oscillations^{16,21,27}, where the frequency of Ca^{2+} transients can be modulated when InsP_3 concentrations are increased^{17,27}. In contrast, type III InsP_3R , by remaining open in the presence of high $[\text{Ca}^{2+}]$ (Fig. 3), initiates a rapid, large, and almost total release of Ca^{2+} from intracellular stores as long as InsP_3 is present (Fig. 4a, d). Thus, type III InsP_3R alone will not support a regenerative response, but its properties make it well suited to initiate intracellular Ca^{2+} signals. In support of this, type III InsP_3R has been localized to the apical region of epithelial cells, the trigger zone from which intracellular Ca^{2+} waves originate^{9–11}. Thus, the single-channel properties of type III InsP_3R are well adapted to its role as the starting gate for Ca^{2+} signals in the cell. □

Methods

Western blots and immunocytochemistry. Immunoblots were probed with antibodies against type I (C-19; custom produced by Research Genetics) or type III (Transduction Lab) InsP_3R . Blots were visualized using ECL (Kirkegaard & Perry). For immunocytochemistry, the same primary antibodies were used to probe for type I and type III InsP_3R in fixed cells. The secondary antibody contained a fluorescent label (FITC) that was observed by confocal microscopy. Nonspecific staining was determined by using only the secondary antibody.

Single-channel recordings. Endoplasmic reticulum vesicles from RIN-5F cells were prepared using the protocol for cerebellum described previously⁷. Vesicles were fused into planar lipid bilayers composed of phosphatidylethanolamine and phosphatidylserine (3:1, w/w; Avanti Polar Lipids) so that the *cis* and *trans* chambers corresponded to the cytosol and lumen of the endoplasmic reticulum respectively. Cytoplasmic bilayer solutions contained 110 mM Tris and 250 mM HEPES (pH 7.35), and luminal solutions contained 55 mM $\text{Ba}(\text{OH})_2$ and 250 mM HEPES (pH 7.35). The *trans* chamber was held at virtual ground and the transmembrane voltage was maintained at 0 mV. Single-channel currents were recorded under voltage-clamp conditions using a patch-clamp amplifier (BC-525B, Warner Instruments) and stored on VHS tape (Instrutech). Data were filtered at 1 kHz and digitized at 4 kHz for computer

analysis using pClamp 6.0.3 (Axon Instruments). For the $I-V$ curve, the membrane potential was clamped at values between 5 and -30 mV. The amplitude of channel openings at each voltage was determined by fitting the data (100–1,600 openings) with a gaussian function. Single-channel conductance was determined by linear regression. An extrapolation to 0 pA to estimate the reversal potential is not included because barium (55 mM $\text{Ba}(\text{OH})_2$) was in the *trans* chamber only, and current through the InsP_3 -gated channel can only flow in one direction from *trans* to *cis*. Current in the opposite direction was not detected. If more positive voltages were included, the current–voltage relationship would begin to curve, and the slope conductance would be underestimated. These are the same experimental conditions as used previously for type I InsP_3R (ref. 7). For the Ca^{2+} dependence curve, calibrated amounts of CaCl_2 were added to the cytoplasmic solution to obtain the desired free Ca^{2+} concentration. We estimated the number of active channels in each bilayer using a statistical model that is dependent on the maximum number of channels observed simultaneously, and then calculated the open probability for a single channel using this value^{2,17}. Open probability data for type I InsP_3R were fitted using the '2- InsP_3 /2- Ca^{2+} ' model¹⁷, whereas open probability data for type III InsP_3R were fitted according to Michaelis–Menten kinetics with $K_m = 0.3$ and $P_{\max} = 5.4$: $P_o = (P_{\max} \cdot [\text{Ca}^{2+}]) / (K_m + [\text{Ca}^{2+}])$.

Cytoplasmic Ca^{2+} measurements. Cytoplasmic Ca^{2+} was measured in single RIN-5F cells or rat hepatocytes using confocal line scanning microscopy²⁸ or in cell populations using ratio spectrofluorimetry²⁹. For single-cell studies, RIN-5F cells or hepatocytes³⁰ were plated onto glass coverslips, incubated at 37°C , and loaded with Fluo-3/AM (6 μM). Coverslips containing the cells were transferred to a perfusion chamber on the stage of a BioRad MRC-600 confocal microscope and observed using a $20\times$ objective³⁰. Increases in $[\text{Ca}^{2+}]$ are expressed as a percentage of baseline fluorescence³⁰. Cells were examined first under control conditions, then in the presence of 0.1 – 100 μM ATP. In selected experiments, RIN-5F cells were stimulated with ATP in Ca^{2+} -free medium containing 1 mM EGTA; similar results were obtained in the presence and absence of extracellular Ca^{2+} . For population studies, cells were loaded with Fura-2/AM (10 μM), then maintained at 37°C in a cuvette. Cells were excited at 340 and 370 nm; fluorescence emission was detected at 485 nm using a PTI DeltaRAM system (these excitation and emission wavelengths were chosen to optimize Fura-2 ratio measurements with this system). Ratios were determined after background subtraction²⁹.

Subcellular Ca^{2+} release. RIN-5F or SKHep1 cells were placed in a perfusion chamber on the stage of a BioRad confocal microscope, then individual cells were pressure-microinjected with a solution containing 1 mM caged InsP_3 , 1 mM Fluo-3, 1 mM HEPES, and 150 mM KCl. Cells were given 5 – 10 min to recover from injection, then InsP_3 was photoreleased using a custom-built system that couples a mercury lamp to a 1 -mm quartz fibreoptic cable through a high-speed shutter and filterwheel while cells were observed using confocal line scanning microscopy²⁸. SKHep1 cells were kindly provided by D. Spray.

Received 23 June; accepted 6 August 1998.

1. Furuchi, T. *et al.* Primary structure and functional expression of the inositol 1,4,5-trisphosphate-binding protein P_{400} . *Nature* **342**, 32–38 (1989).
2. Mignery, G., Sudhof, T. C., Takei, K. & De Camilli, P. Putative receptor for inositol 1,4,5-trisphosphate similar to ryanodine receptor. *Nature* **342**, 192–195 (1989).
3. Sudhof, T. C., Newton, C. L., Archer, B. T., Ushkaryov, Y. A. & Mignery, G. A. Structure of a novel InsP_3 receptor. *EMBO J.* **10**, 3199–3206 (1991).
4. Blondel, O., Takeda, J., Janssen, H., Seino, S. & Bell, G. I. Sequence and functional characterization of a third inositol trisphosphate receptor subtype, $\text{IP}_3\text{R}-3$, expressed in pancreatic islets, gastrointestinal tract, and other tissues. *J. Biol. Chem.* **268**, 11356–11363 (1993).
5. Maranto, A. R. Primary structure, ligand binding, and localization of the human type 3 inositol 1,4,5-trisphosphate receptor expressed in intestinal epithelium. *J. Biol. Chem.* **269**, 1222–1230 (1994).
6. Morgan, J. M., De Smedt, H. & Gillespie, J. I. Identification of three isoforms of the InsP_3 receptor in human myometrial smooth muscle. *Pflügers Arch.* **431**, 697–705 (1996).
7. Bezprozvanny, I. & Ehrlich, B. Inositol (1,4,5)-trisphosphate gated Ca channels from canine cerebellum: divalent cation conduction properties and regulation by intraluminal Ca . *J. Gen. Physiol.* **104**, 821–856 (1994).
8. Perez, P. J., Ramos-Franco, J., Fill, M. & Mignery, G. A. Identification and functional reconstitution of the type 2 inositol 1,4,5-trisphosphate receptor from ventricular cardiac myocytes. *J. Biol. Chem.* **272**, 23961–23969 (1997).
9. Nathanson, M. H., Fallon, M. B., Padfield, P. J. & Maranto, A. R. Localization of the type 3 inositol 1,4,5-trisphosphate receptor in the Ca^{2+} wave trigger zone of pancreatic acinar cells. *J. Biol. Chem.* **269**, 4693–4696 (1994).
10. Thorn, P., Lawrie, A., Smith, P., Gallacher, D. & Petersen, O. H. Local and global cytosolic Ca^{2+} oscillations in exocrine cells evoked by agonists and inositol trisphosphate. *Cell* **74**, 661–668 (1993).
11. Kasai, H., Li, Y. & Miyashita, Y. Subcellular distribution of Ca^{2+} release channels underlying Ca^{2+} waves and oscillations in exocrine pancreas. *Cell* **74**, 669–677 (1993).
12. Yonishima, H., Miyawaki, A., Takayuki, M., Teiichi, F. & Mikoshiba, K. Ca^{2+} differentially regulates

- the ligand-affinity states of the type 1 and type 3 inositol 1,4,5-trisphosphate receptors. *Biochem. J.* **322**, 591–596 (1997).
13. Cardy, T. J. A., Traynor, D. & Taylor, C. W. Differential regulation of types-1 and -3 inositol trisphosphate receptors by cytosolic Ca^{2+} . *Biochem. J.* **328**, 785–793 (1997).
14. Wojcikiewicz, R. J. H. Type I,II,III inositol 1,4,5-trisphosphate receptors are unequally susceptible to down-regulation and are expressed in markedly different proportions in different cell types. *J. Biol. Chem.* **270**, 11678–11683 (1995).
15. DeLisle, S. *et al.* Expression of inositol 1,4,5-trisphosphate receptors changes the Ca signal in *Xenopus* oocytes. *Am. J. Physiol.* **270**, C1255–C1261 (1996).
16. Bezprozvanny, I., Watras, J. & Ehrlich, B. E. Bell-shaped calcium-response curves of $\text{Ins}(1,4,5)\text{P}_3$ - and calcium-gated channels from endoplasmic reticulum of cerebellum. *Nature* **351**, 751–754 (1991).
17. Kaftan, E. J., Ehrlich, B. E. & Watras, J. Inositol 1,4,5-trisphosphate (InsP_3) and calcium interact to increase the dynamic range of InsP_3 receptor-dependent calcium signaling. *J. Gen. Physiol.* **110**, 529–538 (1997).
18. Thomas, A. P., Bird, G. S., Hajnoczky, G., Robb-Gaspers, L. D. & Putney, J. W. Spatial and temporal aspects of cellular calcium signaling. *FASEB J.* **10**, 1505–1517 (1996).
19. Dubyak, G. R. & El-Moatassim, C. Signal transduction via P_2 -purinergic receptors for extracellular ATP and other nucleotides. *Am. J. Physiol. Cell Physiol.* **265**, C577–C606 (1993).
20. Cao, D., Lin, G., Westphale, E., Beyer, E. & Steinberg, T. Mechanisms for the coordination of intercellular calcium signaling in insulin-secreting cells. *J. Cell Sci.* **110**, 497–504 (1997).
21. Dixon, C., Woods, N., Cuthbertson, K. & Cobbold, P. Evidence for two Ca^{2+} -mobilizing purinoceptors on rat hepatocytes. *Biochem. J.* **269**, 499–502 (1990).
22. Bootman, M., Berridge, M. & Lipp, P. Cooking with calcium: the recipes for composing global signals from elementary events. *Cell* **91**, 367–373 (1997).
23. Parker, I. & Yao, Y. Ca^{2+} transients associated with openings of inositol trisphosphate-gated channels in *Xenopus* oocytes. *J. Physiol.* **49**, 663–668 (1996).
24. Lipp, P., Thomas, D., Berridge, M. & Bootman, M. Nuclear calcium signalling by individual cytoplasmic calcium puffs. *EMBO J.* **16**, 7166–7173 (1997).
25. Glennon, M. C., Bird, G. S. J., Kwan, C.-Y. & Putney, J. W. Actions of vasopressin and the Ca^{2+} -ATPase inhibitor, thapsigargin, on Ca^{2+} signaling in hepatocytes. *J. Biol. Chem.* **267**, 8230–8233 (1992).
26. Cameron, A. *et al.* Calcineurin associated with the inositol 1,4,5-trisphosphate receptor–FKBP12 complex modulates Ca^{2+} flux. *Cell* **83**, 463–472 (1995).
27. Petersen, C. C., Toescu, E. C., Potter, B. V. L. & Petersen, O. H. Inositol triphosphate produces different patterns of cytoplasmic Ca^{2+} spiking depending on its concentration. *FEBS Lett.* **293**, 179–182 (1991).
28. Nathanson, M. H., Padfield, P. J., O'Sullivan, A. J., Burgstahler, A. D. & Jamieson, J. D. Mechanism of Ca^{2+} wave propagation in pancreatic acinar cells. *J. Biol. Chem.* **267**, 18118–18121 (1992).
29. Grynkiewicz, G., Poenie, M. & Tsien, R. A new generation of Ca^{2+} indicators with greatly improved fluorescence properties. *J. Biol. Chem.* **260**, 3440–3450 (1985).
30. Schlosser, S. F., Burgstahler, A. D. & Nathanson, M. H. Isolated rat hepatocytes can signal to other hepatocytes and bile duct cells by release of nucleotides. *Proc. Natl. Acad. Sci. USA* **93**, 9948–9953 (1996).

Acknowledgements. We thank E. Kaftan for all his assistance during the execution of these experiments, J. Putney for suggesting RIN-5F cells as source of type III InsP_3R , and J. Brown for comments on the manuscript. This work was supported by NIH grants (to B.E.E. and M.H.N.), an Established Investigator Grant from the American Heart Association (to M.H.N.), and by the Yale Liver Center.

Correspondence and requests for materials should be addressed to R.E.H. at Department of Pharmacology, Yale University, 333 Cedar Street, New Haven, Connecticut 06520-8066, USA (e-mail: rob@hermen.med.yale.edu).

Both Rb/p16^{INK4a} inactivation and telomerase activity are required to immortalize human epithelial cells

Tohru Kiyono*, Scott A. Foster, Jenn I. Koop, James K. McDougall, Denise A. Galloway & Aloysius J. Klingelhutz

Cancer Biology Program, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., CL-015, PO Box 19024, Seattle, Washington 98109-1024, USA

Normal human cells undergo a limited number of divisions in culture and enter a non-dividing state called replicative senescence¹. Senescence is accompanied by several changes, including an increase in inhibitors of cyclin-dependent kinases^{2,3} and telomere shortening⁴. The mechanisms by which viral oncogenes reverse these processes are not fully understood, although a general requirement for oncoproteins such as human papillomavirus E6 and E7 has suggested that the p53 and Rb pathways are targeted. Expression of the catalytic component of telomerase, hTERT, alone significantly extends the lifespan of human fibroblasts⁵. Here we show that telomerase activity is not sufficient for immortalization of human keratinocyte or mammary epithelial

* On leave from the Laboratory of Viral Oncology, Research Institute, Aichi Cancer Center, Nagoya 464, Japan.

cells: we find that neither addition of hTERT nor induction of telomerase activity by E6, both of which are active in maintaining telomere length, results in immortalization. Inactivation of the Rb/p16 pathway by E7 or downregulation of p16 expression, in combination with telomerase activity, however, is able to immortalize epithelial cells efficiently. Elimination of p53 and of the DNA-damage-induced G1 checkpoint is not necessary for immortalization, neither is elimination of p19ARF.

Primary keratinocyte or fibroblast cells derived from human foreskin (HFKs and HFFs) proliferate in culture for a finite, cell-type-specific number of population doublings (PD). In models of cellular senescence, the number of PD at which normal cells senesce has been referred to as mortality stage 1 (M1)⁶. Either human papilloma virus (HPV) E6, which targets p53 for degradation, or E7, which inactivates the retinoblastoma protein Rb, can extend the lifespan of cells past M1, although there is a significant extension of proliferative capacity and the likelihood of establishing immortalized cell lines is greatly increased by the combination of E6 plus E7 (refs 7–9). Human mammary epithelial cells (HMEC), however, appear to senesce in two stages: an initial growth arrest, designated mortality stage 0 (M0), occurs after about 20 PD and is apparently controlled by the Rb/p16 pathway, because M0 can either be prevented by E7 (ref. 10), or a few cells escape spontaneously with reduced p16^{INK4A} expression due to promoter methylation¹¹; the subsequent M1 growth arrest can be bypassed by E6 (refs 12, 13). In some cases, even the combination of viral oncoproteins seems to be insufficient for immortalization, as E6/E7-expressing cells will enter a period of 'crisis' in culture, also called mortality stage 2 (M2), from which only a subpopulation of cells will emerge. The severity of this crisis is cell-type-specific, with growth of HFK or HMEC cultures often being only minimally slowed, whereas in HFFs it is much more pronounced; HFFs are therefore more difficult to immortalize¹⁴. These results indicate that immortalization involves inactivation of several pathways. It has recently been proposed that telomere loss is the sole factor to govern senescence, with hTERT being sufficient to immortalize two normal cell strains, BJ fibroblasts and RPE-340 retinal pigment epithelial cells⁵. We now examine the contribution of hTERT, E6 and E7 to immortalization of primary epithelial and fibroblast cultures in order to determine whether constitutive expression of telomerase alone can cause immortalization.

HFK, HMEC and HFF cultures were infected with retroviral constructs that expressed the hTERT open reading frame¹⁵, either alone or in combination with E7. Both bulk populations and clones were passaged in culture to determine lifespan; the results are summarized in Table 1. LXS (vector)-HFF clones senesced at PD 22–34 after cloning, whereas all hTERT-HFF clones showed a longer lifespan to PD 100 after cloning and are still proliferating, confirming that hTERT extends the lifespan of fibroblasts⁵. However, expression of hTERT did not increase the lifespan of pooled or clonal populations of HFKs or of pre-M0 HMEC. HFK clones containing vector or hTERT alone senesced irreversibly between PD 2 and PD 10 after cloning; one clone went through a severe crisis lasting 7 weeks and then continued to proliferate. In contrast, the lifespan of all seven hTERT/E7-HFK clones was significantly extended and without crisis (>PD 70) and these clones appear to be immortal; the E7-HFK clones had an extended lifespan but senesced at PD 10–20. Likewise, introduction of hTERT alone into pre-M0 HMEC cells did not prevent the M0 growth arrest, whereas introduction of hTERT into post-M0 HMEC (lacking p16 expression), or a combination of E7 and E6 in pre-M0 HMEC, resulted in populations that proliferated without further crisis in culture and have exceeded PD100.

We next tested whether the inability of hTERT to immortalize epithelial cells was related to differences in telomerase induction or telomere maintenance. Infection with hTERT activated telomerase efficiently in HFFs in both normal and E7-expressing HFKs

(Fig. 1a), and in HMEC (Fig. 1b, lane 1). In addition to the pooled populations, 10 of 11 hTERT-HFK clones, 7 of 10 hTERT/E7-HFK clones and 6 of 6 hTERT-HFF clones had telomerase activity, whereas vector clones, with or without E7, did not (data not shown). E6 consistently induced telomerase activity in HFKs and HMECs, but not in fibroblasts, as expected^{16,17}; however, hTERT was significantly more active than E6 (Fig. 1b) and clones of E6-HFKs varied considerably in their telomerase activity (Fig. 1c) for reasons not understood. To demonstrate that this telomerase activity was biologically relevant, we measured telomere length. Extension of telomere length in hTERT-expressing HFKs was apparent within a few PD (Fig. 1d). We were only able to confirm that telomere length was maintained after extensive passaging in one hTERT-HFK population (Fig. 1d, lane 8) because the hTERT-HFK cells senesced. By the time clones of E7-expressing HFKs could be analysed, telomeres had shortened considerably (Fig. 1d, lanes 9–11), although introduction of hTERT into E7 HFKs was sufficient to extend them. As anticipated, hTERT prevented telomere shortening in the HFF clones (Fig. 1e). Although E6 induced telomerase activity in epithelial cells, telomere length was not maintained in pooled populations (results not shown), in agreement with previous studies^{18,19}. We selected E6-HFK clones with either high or low telomerase activity (Fig. 1c) and examined telomere length at a single point in their limited lifespan (Fig. 1f). Clones with high telomerase activity had longer average telomere length than those with less activity. We also compared E7-HFK clones with E6/E7 HFKs because E7 extended the lifespan sufficiently to allow telomere lengths to be compared at different PD but did not induce telomerase activity¹⁶ (Fig. 1a, lane 7, and Fig. 1b, lane 4). Telomeres were shortened in three telomerase-negative E7-HFK clones over 12 PD, but their length was maintained in three telomerase-positive E6/E7-HFK clones over the same period (Fig. 1g). We conclude that E6 induction of telomerase activity is sufficient to maintain telomere length in some clones. Although E6 populations show telomere shortening, those clones that maintain telomere length and have E7, continue to proliferate; however, we cannot exclude the possibility that E7 contributes to telomere maintenance in E6-expressing cells.

Table 1 Immortalization by hTERT, E6 and E7

Cell-type/gene	TRAP	MT*	p16/Rb	p53	Immortal
HFK	–	–	+	+	–
E6	+	+/-†	+	–	–
E7	–	–	–	+/-‡	–
E6 + E7	+	+/-†	–	–	+
hTERT	++	+	+	+	–
E7 + hTERT	++	+	–	+/-‡	+
HMEC-pre M0	–	–	+	+	–
E6	+	NT	+	–	–
E7	–	NT	–	+/-‡	–
E6 + E7	+	NT	–	–	+
hTERT	++	NT	+	+	–
HMEC-post M0	–	–	–	+	–
E6	+	+/-†	–	–	+
E6 – Δ146–151	+	NT	–	–	+
E6 – Δ140–151	–	NT	–	–	–
E6 – Δ123–127	–	NT	–	+	–
E6 – Δ118–122	–	NT	–	–	–
E6 – Δ9–13	–	NT	–	+	–
E6 – 8S9A10T	+	NT	–	+	+
E7	–	NT	–	+/-‡	–
E6 + E7	+	NT	–	–	+
hTERT	+	NT	–	+	+
HFFs	–	–	+	+	–
E6	–	–	+	–	–
E7	–	–	–	+	–
E6 + E7	–	–	–	–	+/-
hTERT	+	+	+	+	+

NT, not tested.

* MT, maintains telomere length.

† Varying levels among clones.

‡ p53 is expressed abundantly but E7 inactivates some functions.

§ Historic data (see also ref. 14).

|| Not efficiently.

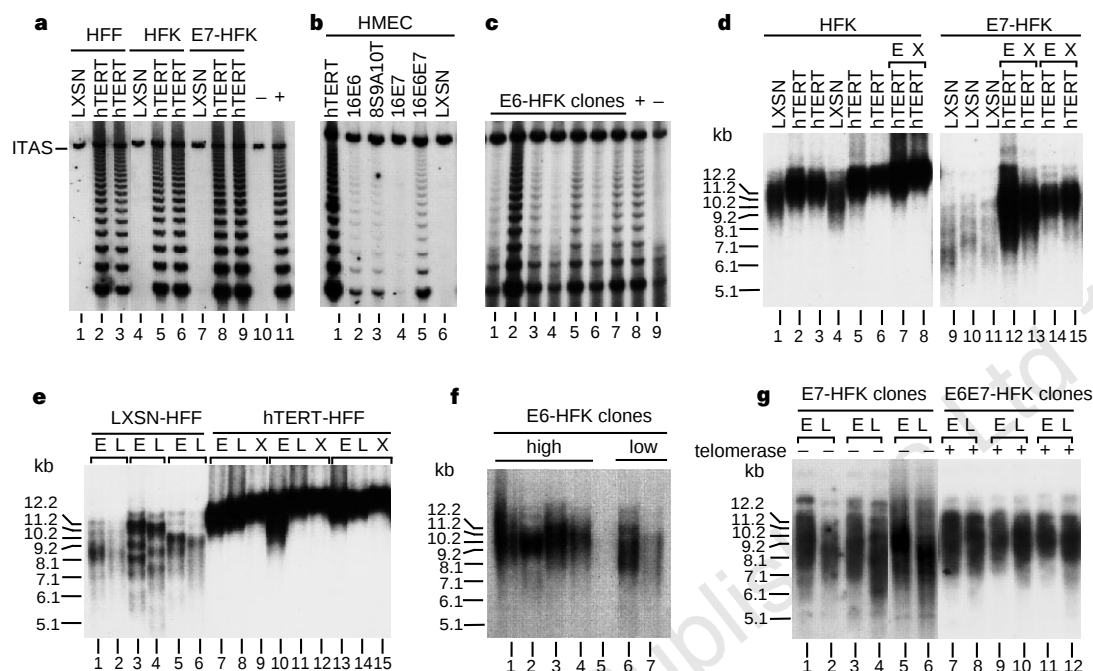


Figure 1 Telomerase activity and telomere length. **a**, Designated cell types were infected with vector LXSN or hTERT as indicated, and telomerase activity was measured by TRAP assay (lanes 1–9); lane 10, lysis buffer alone; lane 11, lysate from a post-crisis HPV 16E6/E7 immortalized HFK. **b**, HMEC infected with designated retroviruses were analysed by TRAP assay. **c**, Individual clones of 16E6-expressing HFKs were assayed for telomerase activity (lanes 1–7); negative and positive controls (lanes 8 and 9) were as for **a**. **d**, Genomic DNA was analysed for telomere length¹⁸ from pools of HFK infected with LXSN or hTERT (lanes 1–3); (lanes 4–7) presenescent (PD 6) clones of HFK infected with LXSN or hTERT; lane 8, the same hTERT clone as in lane 7, which survived crisis and had an extended (X) lifespan (PD 20); lanes 9–11, presenescent LXSN-infected E7-HFK clones at PD

6; lanes 12–15, early (PD 6) and extended (PD 28) passage hTERT expressing E7-HFK clones. **e**, Telomere maintenance was examined in clones of LXSN infected fibroblasts at early (E) PD4, and later (L) PD18 passage (lanes 1–6); lanes 7–15, clones of hTERT expressing fibroblasts at early (PD4), later (PD18) and extended passage (PD50). **f**, Telomere length was analysed in HPV 16E6 expressing HFK clones (PD8) with high telomerase activity (lanes 1–4); lane 5, skipped; and in E6-HFK clones (PD8) with low or no telomerase activity (lanes 6, 7). **g**, Telomere maintenance was analysed: there were three telomerase-negative HPV 16E7-expressing clones at early (PD4) and later (PD16) passages (lanes 1–6), and three telomerase-positive clones that expressed both HPV 16E6 and 16E7 at the same PD as the E7 alone clones (lanes 7–12).

In normal HFKs, the amount of p16 increases at later PD and phosphorylation of Rb is reduced (Fig. 2, lanes 1, 2). Comparison of lanes 1 and 6 (Fig. 2) shows that by the time HFK clones were established, they had undergone enough PD for p16 to have increased. E6- and hTERT-HFKs that senesced at M1 also had raised p16 and reduced Rb phosphorylation (results not shown). Consistent with previous studies, E7 expression resulted in high levels of p16 (ref. 20), and reduced Rb²¹, although some phosphorylated Rb remained (Fig. 2; lanes 3, 5, 8, 9, 14 and 15). Several lines of evidence point to inactivation of the Rb/p16 pathway as being critical for immortalization of epithelial cells. First, E7 cooperates with hTERT or E6 in immortalization, and we previously showed that E7 proteins impaired for Rb binding did not bypass the M0 stage of HMEC¹¹. Second, the spontaneous escape of HMEC from M0 involves downregulation of p16^{INK4a} expression by promoter methylation, resulting in hyperphosphorylation of Rb¹¹ (Fig. 2, lanes 11–13). p16 must have been specifically targeted as p19ARF was not downregulated (Fig. 2, lanes 21–26), and there was no methylation of the p15^{INK4B} promoter¹¹. Even at PD > 100, HMEC continued to express raised amounts of p19ARF (results not shown). Third, p16 was also reduced in the only E6-HFK population to survive extended passage (Fig. 2, lane 4) and in the only hTERT-HFK clone to survive crisis (lane 7); p16 increased progressively in hTERT-HFF cultures (lanes 16–19), which is characteristic of senescent cultures, although Rb phosphorylation was not reduced. There was strikingly less p16 even in PD140 fibroblasts (Fig. 2, lane 19), compared with late-passage (PD 7) epithelial cells (lane 20). The continued phosphorylation of Rb, despite increasing p16, could reflect a disruption of the Rb pathway; alternatively, the low levels of p16 might be insufficient to limit proliferation of

fibroblasts. It is possible that, despite the extended lifespan of hTERT-HFFs, they might eventually senesce when the level of p16 becomes inhibitory.

We next investigated which activity of HPV 16 E6 oncoprotein is involved in immortalization: its degradation of p53 (ref. 22), its binding to the human orthologue of the DLG tumour suppressor²³, and its induction of telomerase activity¹⁶ could all be relevant. Retroviruses expressing wild-type E6 with these properties, or mutated E6 constructs were tested for their ability to immortalize two post-M0 HMEC cultures that had lost expression of p16^{INK4a} as a result of promoter methylation. We found that neither inactivation of p53 nor hDLG binding were required to bypass M1, but that induction of telomerase activity was correlated with immortalization (Table 1). For example, 16E6-8S9A10T, which did not target p53 for degradation (Fig. 3) but induced TRAP activity (Fig. 1b), immortalized cells as efficiently as wild-type 16E6 without any significant crisis. These cells accumulated p53 and p21 and showed an intact G1 checkpoint in response to DNA damage, even after they were immortalized (Fig. 3). 16E6-Δ146–151, which lacks the hDLG-binding motif but induces telomerase, also allowed cells to bypass M1. Conversely, 16E6-Δ140–151, which eliminated p53 but did not induce telomerase, failed to immortalize HMEC3 cells and immortalized HMEC4 cells only after a very extended crisis. Previous results have suggested that p53 degradation is necessary for immortalization of HMEC by E6 (ref. 24). Our assay distinguishes between oncoproteins such as 16E6-8S9A10T that bypass M1, and those such as 16E6Δ118–122 and 16E6Δ140–151 that result in a prolonged crisis from which few cells survive. These latter mutants may still retain a weak ability to activate telomerase, or loss of p53 may allow selection of cells that have undergone other changes. It

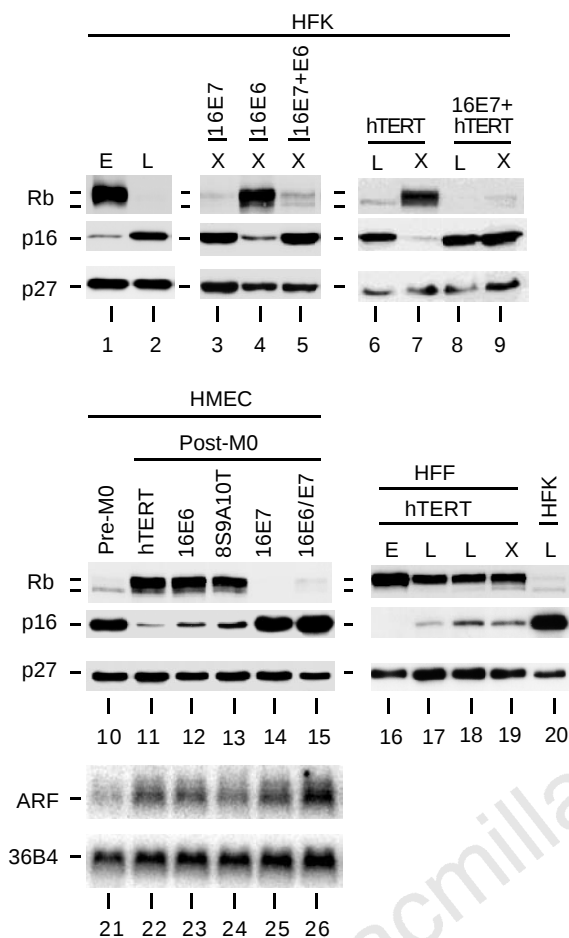


Figure 2 Inactivation of the Rb/p16^{INK4A} pathway. Western blot analysis for Rb, p16^{INK4A} and p27, and northern blot analysis for p19ARF (with 36B4 as a loading control) are shown. HFKs were either uninfected (lanes 1, 2) or transduced with the indicated retroviruses (lanes 3–9) at early passage and collected at early (E), late (L), or extended (X) passage. HMEC were transduced before M0 and were collected either just before reaching M0 (lane 10; vector-infected) or after M0 but before M1 (lanes 11–15 and 21–26). HFFs were transduced with hTERT at early passage and were collected at the passages indicated (lanes 16–19). HFKs at late passage (lane 20) were included on the same gel for comparison with HFFs. Lanes 1–5 and 10–26, pooled populations; lanes 6–9, cloned cells.

will be of interest to see how E6 contributes to immortalization of HFFs. Whereas E6 induces telomerase in epithelial cells, in fibroblast immortalization it may promote a genetic instability that allows selection of cells whose telomerase has been activated by other means, which may explain the greater propensity for epithelial cells to become immortalized by E6/E7 in comparison with fibroblasts. Immortalization of mouse embryo fibroblasts is usually accompanied by loss of either ARF or p53 (ref. 25), but this does not seem to limit proliferation of epithelial cells as immortalized HMEC continue to express both genes. We find that ARF expression is raised in post-M0 and immortalized cells relative to pre-M0 HMEC; however, there was no relation between ARF and p53 expression.

It is also unclear how E6 induces telomerase in epithelial cells. E6 has been reported to increase Myc protein through an undefined post-transcriptional mechanism, and Myc increases hTERT mRNA expression¹⁷; paradoxically, E6 has also been reported to target Myc for degradation²⁶. We have noted (unpublished results) that all E6 proteins that immortalize HMEC retain binding activity to E6AP, a ubiquitin ligase necessary for p53 ubiquitination²⁷. E6 may induce the degradation of cellular proteins other than p53 in an E6AP-

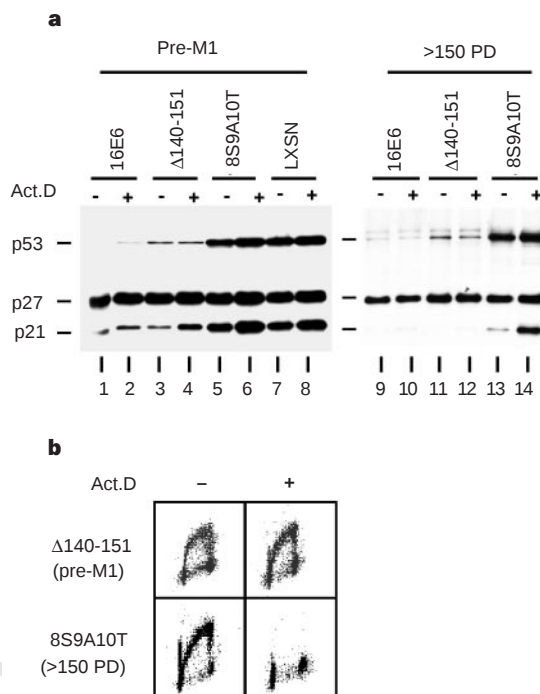


Figure 3 E6 immortalization does not require inactivation of p53. **a**, Western blots for p53, p27 and p21 in cells treated with (+) or without (–) actinomycin D for 24 h. Post-M0 HMEC were transduced with the indicated gene or the backbone vector, LXSN. Cells were used at the indicated stage: pre-M1 (lanes 1–8) or PD >150 (lanes 10–14). **b**, Cell cycle analysed by BrdU/propidium iodide flow cytometry. The same cell populations as in **a** were incubated in the presence of BrdU after treatment with (+) or without (–) actinomycin D.

dependent manner and these proteins might be involved in telomerase repression.

Our results indicate that telomerase activity alone is not sufficient to immortalize two types of primary epithelial cell and that inactivation of the Rb/p16 pathway is a second critical step. In studies using hTERT alone to extend the lifespan of the cultured epithelial cell strains RPE-340 (ref. 5) and HMEC-spiral K (ref. 17), the status of the Rb/p16 pathway was not examined. We have found that HMEC cease to express p16 in culture, and that commercially available HMEC are post-M0 (ref. 10). The sufficiency of hTERT for immortalization of fibroblasts probably reflects cell-type-specific differences in the regulation of senescence, but some questions remain. We found that hTERT extended the lifespan of primary HFF by at least 70 PD, although we cannot rule out the possibility that p16 or some other regulator will eventually inhibit proliferation, or that some feature of the Rb pathway has been disrupted. BJ fibroblasts⁵, but not IMR-90 fibroblasts¹⁷, also had their lifespans significantly extended, indicating that either hTERT levels or some other pathway could be limiting the proliferation of the IMR-90 cells. Our finding that telomerase activation is not enough to immortalize some cell types provides new targets for intervention in oncogenesis and senescence. □

Methods

Retroviral construction and infection. The hTERT retroviral construct was made by removing an *EcoRI* fragment containing full-length hTERT cDNA¹⁵ from pGRN121 (a gift from Geron Corporation) and inserting it into pLXSN. The HPV16 E6, E7, E6/E7 and mutated E6 retroviruses, construction of PG-13 producer cell lines, and infection protocols have been described⁹. For combinations of retroviral infections, cells were first transduced with 16E7 using a hygromycin-selectable retrovirus and then with E6 or hTERT using a neomycin-selectable retrovirus.

Cell culture and growth-arrest assays. Human foreskin keratinocytes, fibroblasts and mammary epithelial cells were isolated as described^{9,13}. HFKs were grown in keratinocyte serum-free medium (K-SFM; Gibco-BRL), HMEC were grown in DFCE-1 (ref. 13) and HFFs were grown in 10% FBS DMEM (Gibco-BRL). Epithelial cells and fibroblasts were selected in 50–100 $\mu\text{g ml}^{-1}$ and 1,000 $\mu\text{g ml}^{-1}$ G418, respectively, or 40 $\mu\text{g ml}^{-1}$ hygromycin B after retroviral infection. HFK and HMEC were infected by PD 4–10, whereas HFF were infected at PD 30. Clones were isolated by cylinder isolation of colonies. HFKs were split 1:4 at early passages, and 1:8 thereafter using trypsin/EDTA. HMEC were fed every other day and passaged at a ratio of 1:5 or 1:10 before cultures were confluent. Fibroblasts were split 1:16. Cells in crisis or senescence were fed 3 to 4 times weekly, split if necessary, and maintained until there was no visible sign of proliferation (usually 2 months). Subconfluent cultures of retrovirally infected HMEC expressing various E6 genes were treated with 2.5 nM actinomycin D and growth arrest assayed as described²¹.

TRAP and telomerase length assays. The telomere repeat amplification protocol (TRAP) assay was performed with several modifications to the original protocol²⁸ in order to control for false positives and inhibitors of PCR amplification. Cells were counted and lysed in CHAPS lysis buffer at a concentration of 6,400 cells per μl buffer and 1 μl was used for the TRAP assay. This number of cells was within the range of exponential amplification of products in the TRAP assay using positive control cells that had high telomerase activity. The TS primer was kinase-labelled with [γ -³²P]ATP and used at 25×10^6 c.p.m. μg^{-1} . The CX-ext primer is a modification of the original CX primer and was designed to prevent primer-dimer interactions and false positives²⁹. An internal ITAS fragment was included to control for the PCR reaction³⁰. PCR cycles were carried out as described and one-eighth of the sample was run on a 7.2% non-denaturing polyacrylamide gel. Relative telomerase signal intensity of the repeat bands was measured by phosphor imaging analysis. Telomere length was assayed as described¹⁸.

Western and northern blot analyses. 20 μg protein from whole-cell extracts was separated by SDS-PAGE and blots were prepared on Immobilon-P membrane (Millipore). The following antibodies were used as probes: for p53 and p21, clones DO-1 and EA-10 (Oncogene Science); for Rb and p16, clones G3-245 and G175-405 (PharMingen); and for p27, K25020 (Transduction Laboratories). Blots were then probed with horseradish-peroxidase-conjugated goat anti-mouse IgG (Jackson ImmunoResearch) and visualized by using the chemiluminescence system (Renaissance; Du Pont NEN). For northern blots, 16 μg total RNA per lane was separated on formaldehyde-agarose gels, transferred to Hybond N membranes (Amersham) and probed with a p16-exon 1 β probe, specific for ARE. Blots were re-probed for acidic ribosomal protein-PO (36B4) as a loading control.

Received 15 May; accepted 27 August 1998.

- Hayflick, L. The limited *in vitro* lifespan of human diploid cell strains. *Exp. Cell Res.* **37**, 614–621 (1965).
- Loughran, O. *et al.* Evidence for the inactivation of multiple replicative lifespan genes in immortal human squamous cell carcinoma keratinocytes. *Oncogene* **14**, 1955–1964 (1997).
- Alcorta, D. A. *et al.* Involvement of the cyclin-dependent kinase inhibitor p16 (INK4a) in replicative senescence of normal human fibroblasts. *Proc. Natl Acad. Sci. USA* **93**, 13742–13747 (1996).
- Allsopp, R. C. *et al.* Telomere length predicts replicative capacity of human fibroblasts. *Proc. Natl Acad. Sci. USA* **89**, 10114–10118 (1992).
- Bodnar, A. G. *et al.* Extension of life-span by introduction of telomerase into normal human cells. *Science* **279**, 349–352 (1998).
- Wright, W. E. & Shay, J. W. The two-stage mechanism controlling cellular senescence and immortalization. *Exp. Gerontol.* **27**, 383–389 (1992).
- Hawley-Nelson, P., Voudsen, K. H., Hubbert, N. L., Lowy, D. R. & Schiller, J. T. HPV 16 E6 and E7 proteins cooperate to immortalize human foreskin keratinocytes. *EMBO J.* **8**, 3905–3910 (1989).
- Munger, K., Phelps, W. C., Bub, V., Howley, P. M. & Schlegel, R. The E6 and E7 genes of the human papillomavirus type 16 together are necessary and sufficient for transformation of primary human keratinocytes. *J. Virol.* **63**, 4417–4421 (1989).
- Halbert, C. L., Demers, G. W. & Galloway, D. A. The E7 gene of human papillomavirus type 16 is sufficient for immortalization of human epithelial cells. *J. Virol.* **65**, 474–478 (1991).
- Foster, S. A. & Galloway, D. A. Human papillomavirus type 16 E7 alleviates a proliferation block in early passage human mammary epithelial cells. *Oncogene* **12**, 1773–1779 (1996).
- Foster, S. A., Wong, D. J., Barrett, M. T. & Galloway, D. A. Inactivation of p16 in human mammary epithelial cells by CpG island methylation. *Mol. Cell. Biol.* **18**, 1793–1801 (1998).
- Shay, J. W., Wright, W. E., Brasikyte, D. & van der Haagen, B. A. E6 of human papillomavirus type 16 can overcome the M1 stage of immortalization in human mammary epithelial cells but not in human fibroblasts. *Oncogene* **8**, 1407–1413 (1993).
- Band, V., Zajchowski, D., Kulesa, V. & Sager, R. Human papilloma virus DNAs immortalize normal human mammary epithelial cells and reduce their growth factor requirements. *Proc. Natl Acad. Sci. USA* **87**, 463–467 (1990).
- Shay, J. W., van der Haagen, B. A., Ying, Y. & Wright, W. E. The frequency of immortalization of human fibroblasts and mammary epithelial cells transfected with SV40 large T-antigen. *Exp. Cell Res.* **209**, 45–52 (1993).

- Nakamura, T. M. *et al.* Telomerase catalytic subunit homologs from fission yeast and human. *Science* **277**, 955–959 (1997).
- Klingelutz, A. J., Foster, S. A. & McDougall, J. K. Telomerase activation by the E6 gene product of human papillomavirus type 16. *Nature* **380**, 79–82 (1996).
- Wang, J., Xie, L. Y., Allan, S., Beach, D. & Hannon, G. J. Myc activates telomerase. *Genes Dev.* **12**, 1769–1774 (1998).
- Klingelutz, A. J., Barber, S. A., Smith, P. P., Dyer, K. & McDougall, J. K. Restoration of telomeres in human papillomavirus-immortalized human anogenital epithelial cells. *Mol. Cell. Biol.* **14**, 961–969 (1994).
- Stoppler, H., Hartmann, D. P., Sherman, L. & Schlegel, R. The human papillomavirus type 16 E6 and E7 oncoproteins dissociate cellular telomerase activity from the maintenance of telomere length. *J. Biol. Chem.* **272**, 13332–13337 (1997).
- Khleif, S. N. *et al.* Inhibition of cyclin D-CDK4/CDK6 activity is associated with an E2F mediated induction of cyclin kinase inhibitor activity. *Proc. Natl Acad. Sci. USA* **93**, 4350–4354 (1996).
- Demers, G. W., Foster, S. A., Halbert, C. L. & Galloway, D. A. Growth arrest by induction of p53 in DNA damaged keratinocytes is bypassed by human papillomavirus 16 E7. *Proc. Natl Acad. Sci. USA* **91**, 4382–4386 (1994).
- Scheffner, M., Werness, B. A., Huibregtse, J. M., Levine, A. J. & Howley, P. M. The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. *Cell* **63**, 1129–1136 (1990).
- Kiyono, T. *et al.* Binding of high risk human papillomavirus E6 oncoproteins to the human homologue of the *Drosophila* discs large tumor suppressor protein. *Proc. Natl Acad. Sci. USA* **94**, 11612–11616 (1997).
- Dalal, S., Gao, Q., Androphy, E. J. & Band, V. Mutational analysis of human papillomavirus type 16 E6 demonstrates that p53 degradation is necessary for immortalization of mammary epithelial cells. *J. Virol.* **70**, 683–688 (1996).
- Kamijo, T. *et al.* Tumor suppression at the mouse INK4A locus mediated by the alternative reading frame product p19ARF. *Cell* **91**, 649–659 (1997).
- Gross-Mesilaty, S. *et al.* Basal and human papillomavirus E6 oncoprotein-induced degradation of Myc proteins by the ubiquitin pathway. *Proc. Natl Acad. Sci. USA* **95**, 8058–8063 (1998).
- Huibregtse, J. M., Scheffner, M. & Howley, P. M. Cloning and expression of the cDNA for E6-AP, a protein that mediates the interaction of the human papillomavirus E6 oncoprotein with p53. *Mol. Cell. Biol.* **13**, 775–784 (1993).
- Kim, N. W. *et al.* Specific association of human telomerase activity with immortal cells and cancer. *Science* **266**, 2011–2015 (1994).
- Krupp, G. *et al.* Molecular basis of artifacts in the detection of telomerase activity and a modified primer for a more robust 'TRAP' assay. *Nucleic Acids Res.* **25**, 919–921 (1997).
- Wright, W. E., Shay, J. W. & Piatyszek, M. A. Modifications of a telomeric repeat amplification protocol (TRAP) result in increased reliability, linearity and sensitivity. *Nucleic Acids Res.* **23**, 3794–3795 (1995).

Acknowledgements. We thank the Geron Corporation for the hTERT cDNA, E. Espling for maintaining stocks of retroviruses, the Image Analysis Laboratory for help with preparing figures, and members of the McDougall and Galloway laboratories for discussion. This work was supported by grants from the NIH to J.K.M. and D.A.G.

Correspondence and requests for materials should be addressed to D.A.G. (e-mail: dgallowa@fhcrc.org).

NMR structure of the histidine kinase domain of the *E. coli* osmosensor EnvZ

Toshiyuki Tanaka*, Soumitra K. Saha†, Chieri Tomomori*, Rieko Ishima‡, Dingjiang Liu‡, Kit I. Tong‡, Heiyoung Park‡, Rinku Dutta†, Ling Qin†, Mark B. Swindells§, Toshimasa Yamazaki¶, Akira M. Ono#, Masatsune Kainosho#, Masayori Inouye† & Mitsuhiro Ikura‡

* Center for Tsukuba Advanced Research Alliance and Institute of Applied Biochemistry, University of Tsukuba, Tsukuba 305-8577, Japan

† Department of Biochemistry, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, New Jersey 08854, USA

‡ Division of Molecular and Structural Biology, Ontario Cancer Institute and Department of Medical Biophysics, University of Toronto, 610 University Avenue, Toronto, Ontario M5G 2M9, Canada

§ Helix Research Institute Inc., Kisarazu 292-0812, Japan

¶ National Institute of Agrobiological Resources, Tsukuba 305-0856, Japan

Department of Chemistry, Tokyo Metropolitan University, Hachioji 192-0397, Japan

Bacteria live in capricious environments, in which they must continuously sense external conditions in order to adjust their shape, motility and physiology¹. The histidine-aspartate phosphorelay signal-transduction system (also known as the two-component system) is important in cellular adaptation to environmental changes in both prokaryotes and lower eukaryotes^{2,3}. In

|| Present address: Inpharmatica Ltd, 60 Charlotte Street, London W1P 2AX, UK

this system, protein histidine kinases function as sensors and signal transducers. The *Escherichia coli* osmosensor, EnvZ, is a transmembrane protein with histidine kinase activity in its cytoplasmic region². The cytoplasmic region contains two functional domains⁴: domain A (residues 223–289) contains the conserved histidine residue (H243), a site of autophosphorylation as well as transphosphorylation to the conserved D55 residue of response regulator OmpR, whereas domain B (residues 290–450) encloses several highly conserved regions (G1, G2, F and N boxes) and is able to phosphorylate H243. Here we present the solution structure of domain B, the catalytic core of EnvZ. This core has a novel protein kinase structure, distinct from the serine/threonine/tyrosine kinase fold, with unanticipated similarities to both heat-shock protein 90 and DNA gyrase B.

The catalytic domain of EnvZ (Fig. 1a, b) assumes an α/β -

sandwich fold: one layer consists of a five-stranded β -sheet (strands B, residues 319–323; D, 356–362; E, 367–373; F, 420–423; and G, 431–436), and the other layer comprises three α -helices (α 1, residues 301–311; α 2, 334–343; and α 4, 410–414). Two adjacent parallel β -strands, B and D, which are connected by helix α 2, exhibit an unusual 'left-handed' connectivity⁵ (see below). The two layers enclose an extensive hydrophobic core, augmented by a small antiparallel β -sheet (strands A, residues 297–299; and C, 330–332) that seals the sandwich at one end. A number of structurally vital hydrophobic residues (Fig. 1c) are highly conserved in homologous histidine kinase domains, as would be expected. A striking feature in the EnvZ fold is the presence of a long polypeptide segment that extends away from the rest of the molecule (Fig. 1b). This segment consists of a short α -helix, α 3 (residues 380–384), followed by a long loop (residues 385–409), which we refer to as the

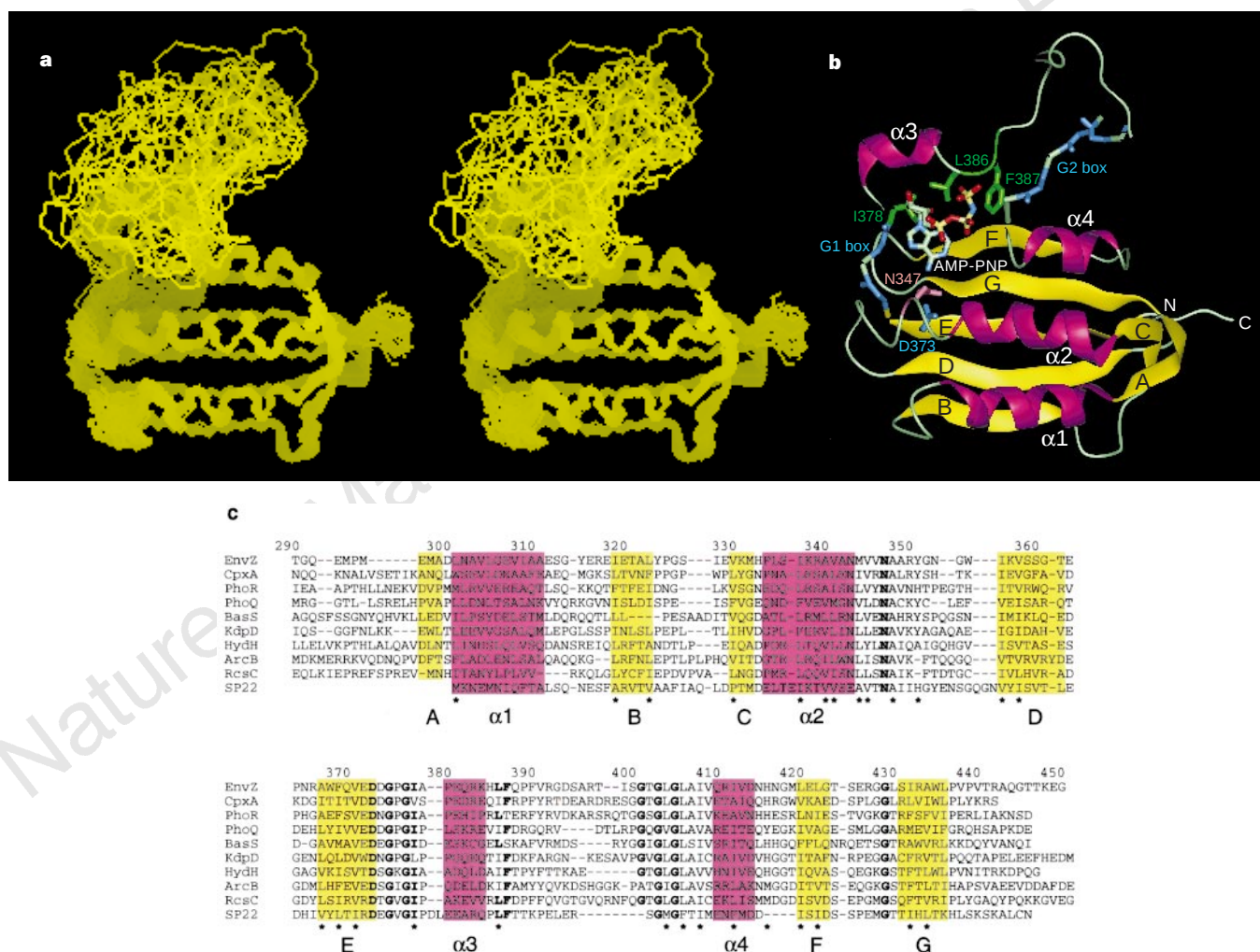


Figure 1 Sequence and structure of the EnvZ catalytic domain. **a**, Stereoview of a best-fit superposition of the backbone (N, C α and C') atoms of the 25 NMR-derived structures of the EnvZ catalytic domain, generated using Insight II (Molecular Simulations). The main-chain atoms of the 25 structures are superimposed against the energy-minimized average structure using residues 297–299, 301–311, 319–323, 330–332, 334–343, 356–362, 367–373, 420–423 and 431–436 (r.m.s.d. of $0.57 \pm 0.12 \text{ \AA}$ for backbone atoms and $0.94 \pm 0.08 \text{ \AA}$ for all heavy atoms). The central loop region is less well-defined because of its flexibility. The N-terminal five and C-terminal ten residues, which are also not well-defined because of the lack of many experimental distance and dihedral angle restraints, are omitted in the figure. **b**, Ribbon drawing of the energy-minimized average structure of the EnvZ catalytic domain. The ATP analogue (AMP-PNP) is shown as a stick model, and the α -helices (in magenta) and β -strands (in yellow) are labelled. The backbone heavy atoms of the glycines in the G1 and G2 boxes (in

blue), and the sidechain heavy atoms of N347 (in pink), D373 (in blue), and I378, L386 and F387 (in green) are also shown as stick models. The N-terminal five and C-terminal ten residues are omitted as in **a**. The N and C termini of the protein are also indicated. The model was generated using QUANTA (Molecular Simulations). **c**, Sequence alignment of the EnvZ catalytic domain with other members of the histidine kinase family. The accession numbers of the SWISS-PROT database for the sequences are P02933 (EnvZ), P08336 (CpxA), P08400 (PhoR), P23837 (PhoQ), P30844 (BasS), P21865 (KdpD), P14377 (HydH), P22763 (ArcB), P14376 (RcsC), and P10728 (SpolIAB; labelled as SP22 in the figure). Secondary structural elements are shown by boxes coloured as in **b**, with their notation indicated below the boxes. Asterisks denote the conserved hydrophobic amino-acid residues. The residues that are important in the ATP binding are shown in blue.

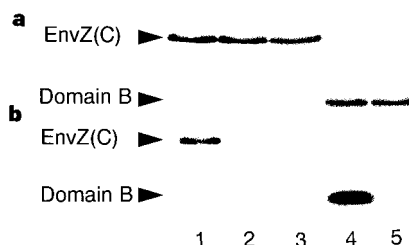


Figure 2 Identification of the ATP-binding domain in EnvZ and the involvement of N347 in ATP binding. **a**, **b**, Ultraviolet-crosslinked protein (see Methods) was visualized through autoradiography (**b**) after 17.5% SDS-PAGE. An identical amount of each protein was used as shown by the CMB-stained SDS-PAGE (**a**). The crosslinked proteins are as follows: lane 1, EnvZ(223-450) (labelled as EnvZ(C) in the figure) with 10 μ Ci [α - 32 P]ATP; lane 2, EnvZ(223-450) with 10 μ Ci [α - 32 P]ATP and 500 μ M non-radioactive ATP; lane 3, EnvZ-N347D(223-450) with 10 μ Ci [α - 32 P]ATP; lane 4, EnvZ(290-450) (labelled as domain B in the figure) with 50 μ Ci [γ - 32 P]ATP; lane 5, EnvZ(290-450) with 50 μ Ci [γ - 32 P]ATP and 1 mM non-radioactive ATP.

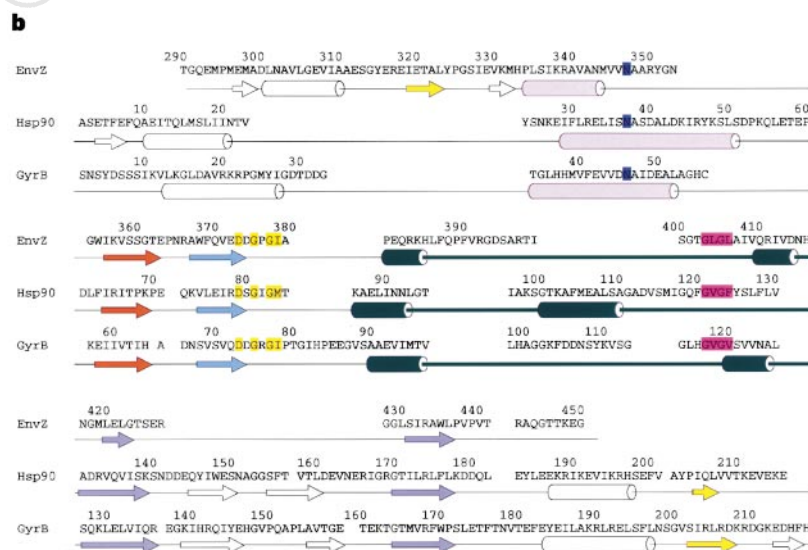
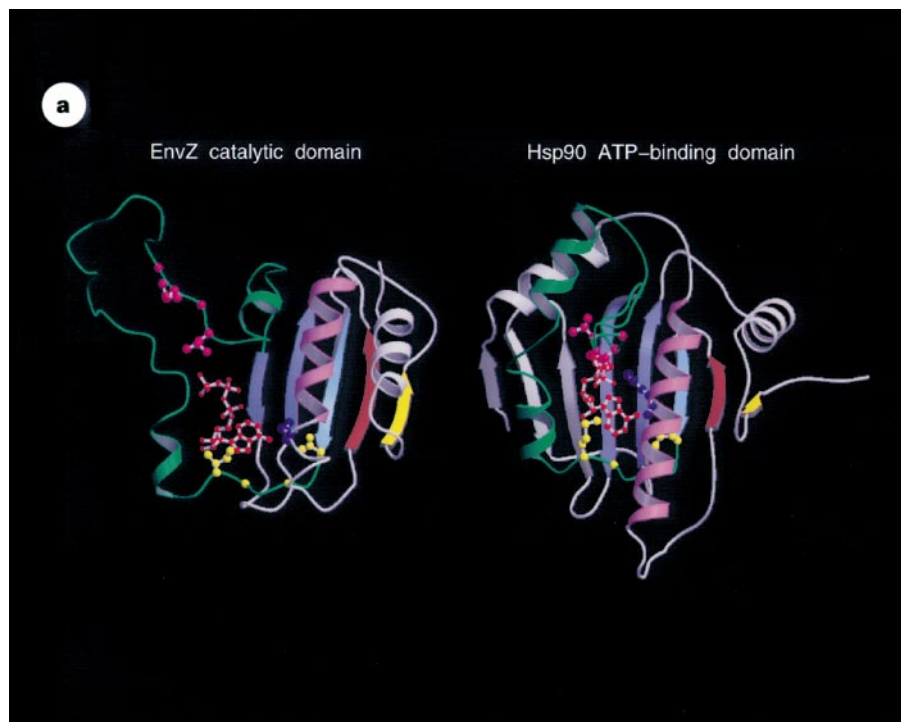


Figure 3 EnvZ and Hsp90 have a similar fold in their ATP-binding regions. **a**, Ribbon representation of the EnvZ catalytic domain and the ATP-binding domain of the molecular chaperone Hsp90 (Protein DataBank accession number 1amw)⁹. For the EnvZ catalytic domain, strand B is shown in yellow, strand D in orange, strand E in light blue, strands F and G in purple, helix α 2 in pink, and the central loop including helices α 3 and α 4 in green. Heavy atoms (except N, C and O) of N347 (in blue), D373, G375, G377 and I378 (in yellow), and G403, L404, G405 and L406 (in magenta) are shown as ball-and-stick models. The corresponding

secondary structural elements and specific residues of Hsp90 are coloured as for EnvZ. AMP-PNP (ADP in Hsp90) is also shown as a ball-and-stick model, in red. The model was generated using MOLSCRIPT²⁰ and Raster3D²¹. **b**, Alignment of the amino-acid sequences of the EnvZ catalytic domain and the ATP-binding domains of Hsp90 and DNA gyrase B (GyrB)^{9,22}, found by the SSAP program²³. α -Helices and β -strands of each structure are indicated as cylinders and arrows, respectively. The colour coding for the secondary structure elements and highlighted residues is as in **a**.

central loop. We detected almost no medium- or long-range nuclear Overhauser effects (NOEs) for the residues in this loop, and the chemical shifts and backbone coupling constants ($^3J_{\text{NH}\alpha}$) were nearly the same as those in a random coil (data not shown). Moreover, ^1H - ^{15}N heteronuclear NOE measurements showed smaller NOE values (0.49 ± 0.10) for the residues in this loop than the values (0.75 ± 0.12) for the structurally well-defined region, further indicating that the region is mobile in solution. The lack of domain A in our NMR structure may be responsible for the high mobility of this central loop. The loop is near to the nucleotide-binding site (see below) and probably interacts with the H243 phosphorylation site and/or its neighbouring residues in domain A.

Like eukaryotic protein kinases, histidine kinases require the presence of ATP and Mg^{2+} for their activity⁶. Non-radioactive ATP competes with [α - ^{32}P]ATP for binding to the whole cytoplasmic domain of EnvZ (residues 223–450) (Fig. 2, lanes 1, 2). This ATP-binding function is retained in domain B (Fig. 2, lane 4); non-radioactive ATP competes with [γ - ^{32}P]ATP for binding to this domain (Fig. 2, lane 5). Our domain B structure contains a non-hydrolysable analogue of ATP, AMP-PNP (β , γ -imidoadenosine-5'-triphosphate), whose location in the structure has been determined on the basis of a dozen intermolecular NOEs observed between sidechain protons of I378/L386/L422 and the adenosine H2, H8, H1', H4' and H5' protons. The AMP-PNP molecule is surrounded by helix α 3 and part of the central loop, and makes extra contacts with strands F and G. The AMP-PNP adenine ring is in close spatial proximity to N347, D373, I378, L386 and F387 (Fig. 1b), which are conserved in members of the histidine kinase family (Fig. 1c). The triphosphate chain is exposed to the protein surface, consistent with the potential for transferring the γ -phosphate to H243 in domain A. In addition, five glycines (G375, G377, G403, G405 and G429) and an asparagine (N347) in the catalytic core are highly conserved and strategically located in the structure (Fig. 1b), indicating their structural and functional significance. Previous mutagenesis studies⁷ showed that the glycine-rich regions, G1 (D373–G377) and G2 (G403–G405), are essential for kinase activity. In addition our structure shows that G375 and G429 are particularly important,

allowing sharp kinks (between strand E and helix α 3 and between strands F and G, respectively) that contour part of the AMP-PNP-binding site.

High deviation of the central loop in the NMR-derived structure, due to high mobility, precludes close examination of the residues that could be involved in the catalysis. However, further information can be obtained from several data base searches and structural comparisons. While the structure of EnvZ domain B was being determined, we noted that detailed Ψ -BLAST searches⁸ indicated a distant similarity of domain B with both heat-shock protein 90 (Hsp90) (ref. 9) and DNA gyrase B¹⁰. Although the Ψ -BLAST scores were low, this potential relationship was particularly interesting, because first, all these proteins must bind ATP to fulfil their functions, and second, the ATP-binding domains of Hsp90 and DNA gyrase B were already known to have similar folds⁹. Upon determination, close examination of these structures indicated that the EnvZ fold is indeed similar to the fold found in the ATP-binding regions of Hsp90 and the DNA gyrase B (Fig. 3). Common to all three families are helix α 2, four strands (D, E, F and G) and the central loop (residues 385–409). Together they enclose N347, the G1 (DxGxG ϕ) motif (residues 373–378), and the G2 (G ϕ G ϕ) motif (residues 403–406) (x represents any amino acid; ϕ represents a hydrophobic amino acid), which occur in all three families and are close to the ATP-binding site of Hsp90 (Protein DataBank codes 1amw and 1am1).

In detail, residue N37 of the Hsp90-ADP complex (corresponding to N347 in EnvZ) binds directly to β -phosphate and Mg^{2+} and indirectly to the adenine base⁹. Expression of an EnvZ mutant protein in which N347 is mutated to aspartate (N347D) results in a phenotype in which ATP-dependent autokinase activity is lost; however, the ability of the mutant protein to phosphorylate OmpR is retained¹¹. The importance of N347 is shown in Fig. 2 (lane 3): ATP binding to EnvZ residues 223–450 is abolished in the N347D mutant. Decreased affinity of the N347D mutant for ATP is evidence that this conserved asparagine is vital for the ATP-dependent autophosphorylation activity of EnvZ. In the structures of both Hsp90 and DNA gyrase, the only highly conserved charged residue, aspartate in the DxGxG ϕ motif (D79 and D73 in Hsp90 and DNA gyrase, respectively), interacts with the N6 atom in the adenine base^{9,10}. The corresponding residue in EnvZ, D373, is also close to the N6 atom of AMP-PNP in our structure (Fig. 1b).

Despite the clear similarity between EnvZ, Hsp90 and DNA gyrase B, EnvZ initially appears to be the outlier of the trio, mainly because of the left-handed $\beta\alpha\beta$ connection between strands B and D. Cases of clear left-handed connectivity are quite rare despite the present availability of more than 1,000 non-homologous structures. It is, therefore, all the more surprising that one such case is found in the carboxy-terminal, DNA-binding domain of *E. coli* DNA gyrase B (Fig. 4). The similarity of the ATP-binding domain, coupled with the presence of a left-handed $\beta\alpha\beta$ connection in both proteins, supports the possibility of a distant evolutionary link between these two protein families.

In an analysis of 11 completed genomes using the EnvZ catalytic domain we identified nearly 150 homologous sequences, using Ψ -BLAST search⁸ with an expectation (*E*) value of 0.01. In a similar search of the non-redundant OWL database we identified over 400 such sequences. These included proteins such as anti-sigma factor¹² (SpoIIB). On the *E. coli* chromosome, genes encoding 32 histidine kinases have been identified¹³, and 17 of these kinases have been characterized biochemically². The sequence of Sln1, an osmosensor in yeast¹⁴, is conserved in all regions corresponding to secondary structural elements in EnvZ, but contains a 120-residue insertion between strands D and E of the EnvZ structure. This indicates that the yeast histidine kinase contains the same fold as that of EnvZ, with a modification between the two strands, presumably resulting in additional functionality.

There are at least 100 examples of histidine–aspartate phosphorelays

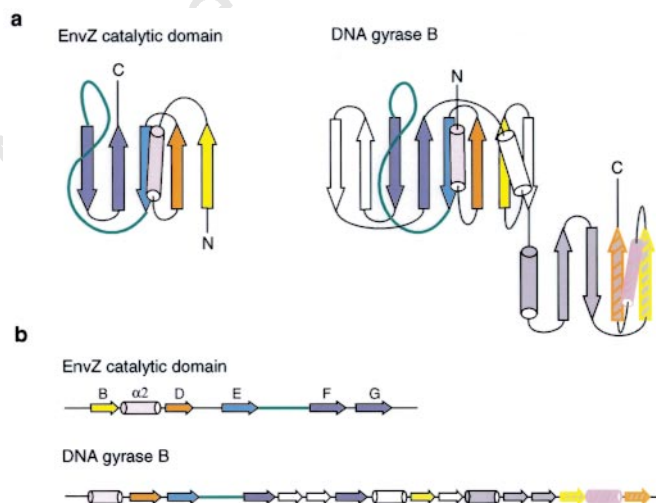


Figure 4 The EnvZ catalytic domain and DNA gyrase B. **a**, Secondary structure; **b**, the linear arrangement of structural elements. This figure emphasizes the recurrence of aligned structural elements involved in the ATP-binding sites (coloured) and the position of the left-handed $\beta\alpha\beta$ motif in each structure (EnvZ, yellow–pink–orange; DNA gyrase B, hashed regions of the same colours). α -Helices and β -strands of each structure are indicated as cylinders and arrows, respectively. The colour coding for the secondary structure elements is the same as in Fig. 3a except for the C-terminal domain of the DNA gyrase B subunit. The N and C termini of each structure are indicated.

in bacteria, and some are implicated in bacterial virulence². For example, a life-threatening bacterium, *Salmonella typhimurium*, possesses the PhoP/PhoQ phosphorelay system which seems to be essential for its virulence in host organisms¹⁵. As all of these phosphorelay systems contain a conserved histidine kinase domain, which has not been found in mammalian cells, histidine kinases are excellent targets for antimicrobial action¹⁶. Members of a family of hydrophobic tyramines have already been identified as histidine kinase inhibitors; the half-maximal inhibitory concentration of these inhibitors is 2–600 μM (ref. 17). Our structure of the *E. coli* EnvZ histidine kinase catalytic domain will provide a vital foundation for the rational design of potent antibiotics that specifically inhibit multiple histidine kinases in this and other microbial species. □

Methods

Sample preparation. Recombinant uniformly ¹⁵N- and ¹³C-labelled EnvZ was expressed in overproducing *E. coli* strain pET11a/BL21(DE3) grown in M9 minimal medium, and was purified as described⁴. NMR samples contained 1.0–1.5 mM uniformly ¹⁵N- or ¹³C/¹⁵N-labelled, or unlabelled, protein in either 95% H₂O/5% ²H₂O or 99.996% ²H₂O containing 20 mM sodium phosphate, 50 mM KCl, 0.5 mM 4-(2-aminoethyl)-benzenesulphonyl fluoride, 50 μM sodium azide and 5 mM MgCl₂, pH 7.0, with 5 mM unlabelled or ¹⁵N/¹³C-labelled AMP-PNP.

NMR spectroscopy. All NMR spectra were acquired at 25 °C on Varian UNITY-plus 500, UNITY-600, and Bruker DMX750 spectrometers. ¹H, ¹³C, and ¹⁵N resonances of the backbone were assigned by analysing four triple-resonance experiments, (HB)CBCA(CO)NNH, HNCACB, (HB)CBCACO(CA)HA, and HNCO. Assignment of the sidechain resonances was done using the three-dimensional HCCH-TOCSY and ¹³C-edited NOESY-HMQC spectra. ³J_{NH α coupling constants were measured from ¹H/¹⁵N HMQC-J spectrum, and slow-exchanging amide protons were identified by recording a series of gradient-enhanced ¹H-¹⁵N HSQC spectra at different time points immediately after the H₂O buffer was changed to a ²H₂O buffer.}

Structure calculation. Structure calculations were done using a restrained molecular dynamics simulated annealing protocol¹⁸ within X-PLOR¹⁹. For structure calculations we used 1,782 interproton distance restraints (comprising 556 intraresidue, 440 sequential, 260 short-range, 507 long-range, 13 protein-ATP analogue, and 10 intra-analogue distances) obtained from heteronuclear three-dimensional NOE spectra. In addition to the NOE-derived distance restraints, 92 distance restraints for 46 hydrogen bonds and 122 dihedral angle restraints were included in the structure calculation. The average root mean square deviation (r.m.s.d.) values from idealized geometry for bonds, angles and impropers are 0.005 Å, 0.61° and 0.39°, respectively. The total and Lennard-Jones potential energies are 516 ± 70 and -134 ± 29 kcal mol⁻¹, respectively (calculated with the use of square-well potentials for the experimental distance empirical energy term with a force constant of 50 kcal mol⁻¹ Å⁻²). None of the structures has violations greater than 0.40 Å (for distance restraints) and 3.0° (for dihedral angle restraints).

Ultraviolet-crosslinking assay. Purified proteins were each incubated in crosslinking buffer (20 mM Tris, pH 7.8, 50 mM KCl and 5% glycerol) containing [α -³²P]ATP (800 Ci mmol⁻¹, 10 mCi ml⁻¹) or [γ -³²P]ATP (3,000 Ci mmol⁻¹, 10 mCi ml⁻¹) in a final volume of 25 μl for 15 min at 4 °C. Proteins were crosslinked by ultraviolet irradiation at 254 nm at a height of 4.5 cm for 5 min on ice using an ultraviolet lamp (Model UVG-45, 115 V, 60 Hz, 0.16 A, UVP Inc., California). Crosslinked protein was visualized by autoradiography after analysis by 17.5% SDS-PAGE. For competition experiments, non-radioactive ATP was added in the reaction mixture.

Received 19 June; accepted 12 August 1998.

- Parkinson, J. S. in *Two-Component Signal Transduction* (eds Hock, J. A. & Silhavy, T. J.) 9–23 (ASM, Washington, DC, 1995).
- Egger, L. A., Park, H. & Inouye, M. Signal transduction via the histidyl-aspartyl phosphorelay. *Genes Cells* **2**, 167–184 (1997).
- Wirgler-Murphy, S. M. & Saito, H. Two-component signal transducers and MAPK cascades. *Trends Biochem. Sci.* **22**, 172–176 (1997).
- Park, H., Saha, S. K. & Inouye, M. Two-domain reconstitution of a functional protein histidine kinase. *Proc. Natl Acad. Sci. USA* **95**, 6728–6732 (1998).
- Branden, C. & Tooze, J. *Introduction to Protein Structure* (Garland, New York, 1991).
- Stock, J. B., Surette, M. G., Levitt, M. & Park, P. in *Two-Component Signal Transduction* (eds Hock, J. A. & Silhavy, T. J.) 25–51 (ASM, Washington, DC, 1995).
- Yang, Y. & Inouye, M. Requirement of both kinase and phosphatase activities of an *Escherichia coli*

- receptor (Tazl) for ligand-dependent signal transduction. *J. Mol. Biol.* **231**, 335–342 (1993).
- Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
- Prodromou, C. et al. Identification and structural characterization of the ATP/ADP-binding site in the Hsp90 molecular chaperone. *Cell* **90**, 65–75 (1997).
- Wigley, D. B. et al. Crystal structure of an N-terminal fragment of the DNA gyrase B protein. *Nature* **351**, 624–629 (1991).
- Dutta, R. & Inouye, M. Reverse phosphotransfer from OmpR to EnvZ in a kinase⁺/phosphatase⁺ mutant of EnvZ (EnvZ-N347D), a bifunctional signal transducer of *Escherichia coli*. *J. Biol. Chem.* **271**, 1424–1429 (1996).
- Min, K.-T. et al. Sigma F, the first compartment-specific transcription factor of *B. subtilis*, is regulated by anti-sigma factor that is also a protein kinase. *Cell* **72**, 735–742 (1993).
- Mizuno, T. Compilation of all genes encoding two-component phosphotransfer signal transducers in the genome of *Escherichia coli*. *DNA Res.* **4**, 161–168 (1997).
- Ota, I. M. & Varshavsky, A. A yeast protein similar to bacterial two-component regulators. *Science* **262**, 566–569 (1993).
- Soncini, F. C. & Groisman, E. A. Two-component regulatory systems can interact to process multiple environmental signals. *J. Bacteriol.* **178**, 6796–6801 (1996).
- Dziejman, M. & Mekalanos, J. J. in *Two-Component Signal Transduction* (eds Hock, J. A. & Silhavy, T. J.) 305–317 (ASM, Washington, DC, 1995).
- Barrett, J. F. et al. Antibacterial agents that inhibit two-component signal transduction systems. *Proc. Natl Acad. Sci. USA* **95**, 5317–5322 (1998).
- Nilges, M., Gronenborn, A. M., Brünger, A. T. & Clore, G. M. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. *Protein Eng.* **2**, 27–38 (1988).
- Brünger, A. T. *X-PLOR Version 3.1: A system for X-ray Crystallography and NMR* (Yale Univ. Press, New Haven, 1992).
- Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
- Merritt, E. A. & Murphy, M. E. P. Raster 3D Version 2.0—a program for photorealistic molecular graphics. *Acta Crystallogr. D* **50**, 869–873 (1994).
- Holdgate, C. A. et al. The entropic penalty of ordered water accounts for weaker binding of the antibiotic novobiocin to a resistant mutant of DNA gyrase: a thermodynamic and crystallographic study. *Biochemistry* **36**, 9663–9673 (1997).
- Orengo, C. A., Brown, N. P. & Taylor, W. R. Fast structure alignment for protein databank searching. *Proteins* **14**, 139–167 (1992).

Acknowledgements. We thank L. Kay for providing NMR pulse sequences, L. Pearl for the Hsp90-ATP coordinates before PDB release and discussions, and S. Bagby for comments on the manuscript. This work was supported by grants from JSPS (to T.T.) from CREST (to M.K.) from the NIH (to M. Inouye), and from HHMI (to M. Ikura). R.I. and D.L. acknowledge HFSP postdoctoral fellowships. M. Ikura is an HHMI International Research Scholar and MRCC Scientist.

Correspondence and requests for materials should be addressed to M. Ikura (e-mail: mikura@oci.utor-onto.ca). Coordinates have been deposited in the Brookhaven Protein Data Bank.

Structure of the haemagglutinin-esterase-fusion glycoprotein of influenza C virus

Peter B. Rosenthal*†, Xiaodong Zhang*†, Frank Formanowski†‡, Wolfgang Fitz†§, Chi-Huey Wong§, Herbert Meier-Ewert‡, John J. Skehel|| & Don C. Wiley¶

* Department of Molecular and Cellular Biology, Harvard University, and
¶ Howard Hughes Medical Institute, 7 Divinity Avenue, Cambridge, Massachusetts 02138, USA

‡ Abteilung für Virologie, Technical University of Munich, Biedersteiner Strasse 29, 80802 Munich 40, Germany

§ Department of Chemistry, The Scripps Research Institute, 10666 North Torrey Pines Road, La Jolla, California 92037, USA

|| National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

The spike glycoproteins of the lipid-enveloped orthomyxoviruses and paramyxoviruses have three functions: to recognize the receptor on the cell surface, to mediate viral fusion with the cell membrane, and to destroy the receptor. In influenza C virus, a single glycoprotein, the haemagglutinin-esterase-fusion (HEF) protein, possesses all three functions (reviewed in ref. 1). In influenza A and B, the first two activities are mediated by haemagglutinin and the third by a second glycoprotein, neuraminidase. Here we report

† Present addresses: MRC Laboratory of Molecular Biology, Hills Road, Cambridge, England CB2 2QH, UK (P.B.R.); Molecular Structure and Function Laboratory, Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London, WC2A 3PX, UK (X.Z.); Cressier, Switzerland (F.F.). Quest International, Bussum, The Netherlands (W.F.).

the crystal structure of the HEF envelope glycoprotein of influenza C virus. We have identified the receptor-binding site and the receptor-destroying enzyme (9-*O*-acylesterase) sites, by using receptor analogues. The receptor-binding domain is structurally similar to the sialic acid-binding domain of influenza A haemagglutinin, but binds 9-*O*-acetylsialic acid. The esterase domain has a structure similar to the esterase from *Streptomyces scabies* and a brain acetylhydrolase^{2,3}. The receptor domain is inserted into a surface loop of the esterase domain and the esterase domain is inserted into a surface loop of the stem. The stem domain is similar to that of influenza A haemagglutinin, except that the triple-stranded, α -helical bundle diverges at both of its ends, and the amino terminus of HEF2, the fusion peptide, is partially exposed. The segregation of HEF's three functions into structurally distinct domains suggests that the entire stem region, including sequences at the amino and carboxy termini of HEF1 which precede the post-translational cleavage site between HEF1 and HEF2, forms an independent fusion domain which is probably derived from an ancestral membrane fusion protein.

The structure of the HEF trimer is shown in Fig. 1. The HEF monomer is composed of three domains: an elongated stem (red in Fig. 2a) active in membrane fusion (F), a receptor-destroying esterase domain (E) (green in Fig. 2a), and a receptor-binding domain (R) (blue in Fig. 2a). Two of these compact domains are made from non-contiguous segments of amino-acid sequence (Fig. 2b): the stem domain F consists of the amino-terminal amino acids 1–40 and carboxy-terminal residues 367–432 of HEF1 and all of HEF2 (labelled F1, F2, F3 in Fig. 2a, b); the esterase domain E consists of HEF1 segments comprising residues 41–150, which precede the receptor domain, and residues 311–366, which follow R (labelled E1, E' and E2 in Fig. 2a, b). The single-segment R domain is inserted into a surface loop of the esterase domain, and the esterase domain is inserted into a surface loop near the top of the stem domain F (Fig. 2c). The R and E domains of HEF are both compact, having their N and C termini within a few angstroms of each other, so that they can be accommodated into a pre-existing

protein at surface loops without disrupting either protein's structure or function.

Alignment of the amino-acid sequences of HEF and HA based on their three-dimensional structures indicates that they have 12% sequence identity (the alignment is available from the authors). Nevertheless, both the overall structure (compare Fig. 2a and d) and the detailed folds of individual segments (Fig. 2a) of HEF and HA are remarkably similar. This is true for the globular R domain, as well as for the highly extended segments F1 and F2 (Fig. 2a) and for the similar helical-hairpin and membrane-proximal five-stranded β -sheet forming HEF2 and HA2 (F3 in Fig. 2a). Two of the segments of the enzyme domain, E1 and E2, are not found in HA (compare Fig. 2b and e), which is consistent with its lack of enzyme activity. If the E' domain of HA is a vestigial fragment of E, then HA may have evolved by the deletion of segments E1 and E2 from an ancestral gene similar to HEF. Segments E1, E', R, E2, and F2 are present, with 30% sequence identity, in the haemagglutinin-esterase (HE) found in coronaviruses⁴. (Data concerning the antigenicity of HEF and the structural relation between HEF and HE will be published elsewhere.)

The R domains of both HEF and HE are similar eight-stranded 'Swiss rolls' of β -sheet (Fig. 2a, d) containing the receptor-binding sites bounded by an α -helix, a loop, and an extended strand (superimposed in Fig. 3b) (except for one residue, Tyr 127 from E'). The structures of the complexes of HEF with two receptor analogues^{5,6}, 9-acetamidodialic acid α -methylglycoside (Fig. 3a) and 9-acetamidodialic acid α -thiomethylmercuryl glycoside⁷, show that sialosides bind similarly to HEF and to HA, although they are recognized by different amino-acid side chains (see ref. 8 and references therein). Sialic acid linkage specificity (α (2,3) vs α -(2,6)) has been attributed to residues near amino acid 226 in HA^{8–10}. The homologous loop (near HEF 270) is truncated in HEF (Fig. 2b), consistent with the lack of linkage specificity of influenza C virus¹¹. The HEF receptor differs from the HA receptor by the addition of an acetyl group at the 9-*O* position of the glycerol side chain. The acetyl methyl group binds in a nonpolar pocket unique to HEF (Fig. 3a).

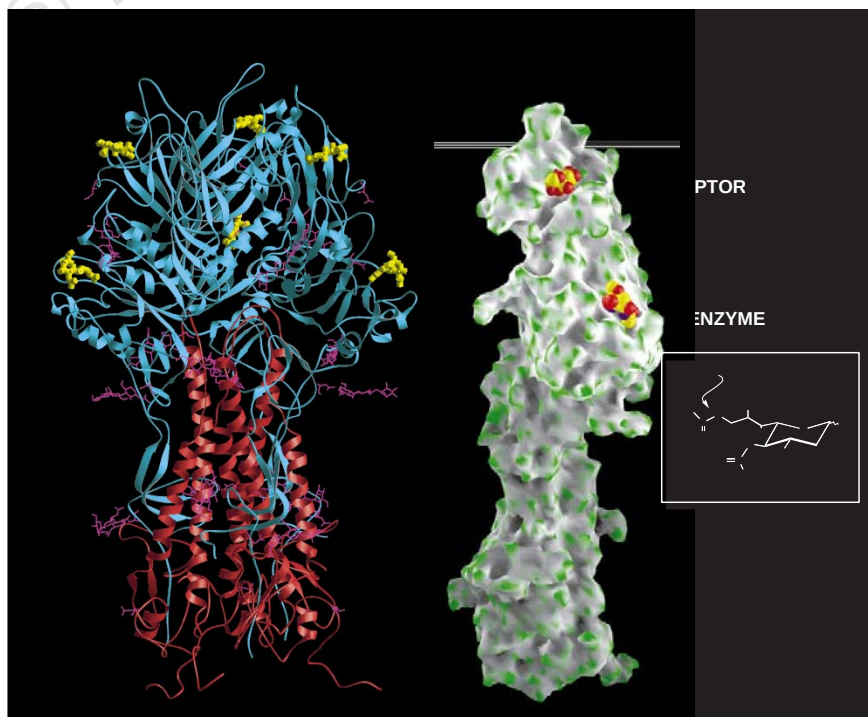


Figure 1 Haemagglutinin-esterase-fusion glycoprotein structure. **a**, The structure of the HEF trimer. HEF1 (blue), HEF2 (red), receptor analogue and enzyme inhibitor ligands, yellow; *N*-linked carbohydrate ball-and-stick (purple). HEF1 is linked to HEF2 by a disulfide bond from Cys 6 of HEF1 to Cys 137 of HEF2. **b**,

Monomer surface of HEF (Grasp²⁷) showing 9-*O*-acetylsialoside receptor binding site (top) and 9-*O*-acylesterase site (bottom). Inset, the esterase removes the acetyl group of 9-*O*-acetylsialic acid (see arrow).

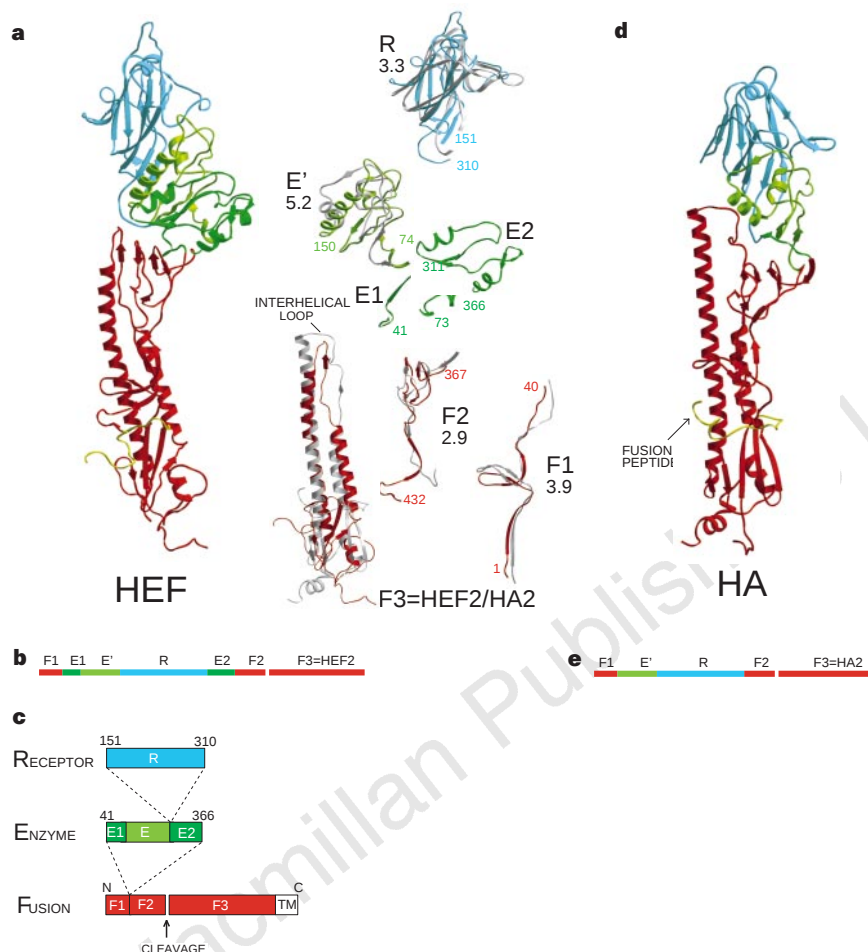


Figure 2 Comparison of HEF and HA monomers. **a**, HEF monomer structure and exploded views of the sequence segments (coloured by domain). Superposition of corresponding HA (X:31 strain; PDB code 1hge) segments (grey) are shown with r.m.s. values below the segment name. Fusion peptides are yellow. **b**, Linear order of the sequence segments in HEF, coloured by domains. Red segments: F1,

F2, F3; green: E1, E2; light green: E'; blue: R. **c**, Topological relationship of the compact domains in HEF. **d**, HA monomer structure with sequence segments coloured by domains. **e**, Linear order of the sequence segments in HA, coloured by domain. Red segments: F1, F2, F3; light green: E'; blue: R.

Complexes of HEF with the two non-hydrolysable receptor analogues described above also show how substrate interacts with the novel 9-*O*-acetyl esterase active site (Fig. 4a). Ser 57 in the catalytic triad (Ser 57 from E1, His 355 and Asp 352 from E2) is positioned for nucleophilic attack on the carbonyl carbon of the 9-*O* acetyl group of the bound sialoside. The carbonyl oxygen points into an 'oxanion hole' formed by the side chain of Asn 117, and the NH groups of Gly 85 and Ser 57 (Fig. 4a). The ligand interactions and the structure of the enzyme site are completely different from those of the receptor-binding site.

A search for proteins with structural similarity to the HEF enzyme domain identified the α 1 subunit of platelet-activating factor acetylhydrolase (Ib) from bovine brain (PAF-AH)² and the esterase from *Streptomyces scabies* (SsEst)¹ (Fig. 4b). All three proteins have a similar topology, despite sharing only 13% sequence identity, and their core residues superimpose (r.m.s. ~ 3 Å), including the central five-stranded β -sheet and long flanking helices (blue in Fig. 4b). When the core folds are superimposed, the catalytic triads (green in Fig. 4b) and oxanion-hole residues of each protein overlap.

Interactions between the α -helices in the stem of the HEF trimer and the packing of the N-terminal fusion peptide of F3 have implications for low-pH-induced conformational change in HEF and the mechanism by which this induces membrane fusion at low pH. The F3 segments of HA and HEF are very similar in structure (Fig. 2a): each monomer has a central α -helix along the threefold axis and a smaller N-terminal helix packed antiparallel on the

outside and connected by an interhelical loop (Fig. 2a). In HEF2, although the central helices interact closely in the middle like HA, they diverge from the trimer axis at both ends (Fig. 5a). At the top, the interhelical loops interpose between the first five turns of the long helices (residues 80–97), where loop residues HEF2 Arg 69 and central helix residues HEF2 Glu 95 form salt bridges and contact an unidentified ion (possibly a sulphate) on the trimer axis. Although HEF2 has a sequence deletion of seven residues in the loop region, this difference would preserve the register of the heptads during the formation of an extended triple-stranded coil in the low-pH conformation, like that in HA¹². Unlike HA, however, the top third of the triple-stranded helical bundle must first make interactions on the trimer axis after the removal of the interhelical loop, before the N-terminal coiled-coil extension can form.

Three tryptophans (HEF2 116) form the last interaction on the trimer axis (Fig. 5a), below which the helices diverge as in HA. Unlike HA, in which residues 2 (Leu) and 3 (Phe) of the HA2 N-terminal 'fusion peptides' interact across the trimer axis, HEF2 residues Val 11 and Leu 12 are closest to the trimer axis but further penetration is blocked by tryptophans at position 116. Residues N-terminal to residue 10 fold back out to the surface of the protein, where aspartic acids at positions 5 and 6 of HEF2 can interact with Arg 29 and Lys 30 of HEF2 and Lys 4 of HEF1. Residues 1–4 of HEF2 appear to be disordered on the surface of the molecule. In HEF, the residues that are functionally analogous to the fusion peptide of HA, namely the buried residues whose exposure would convert a soluble

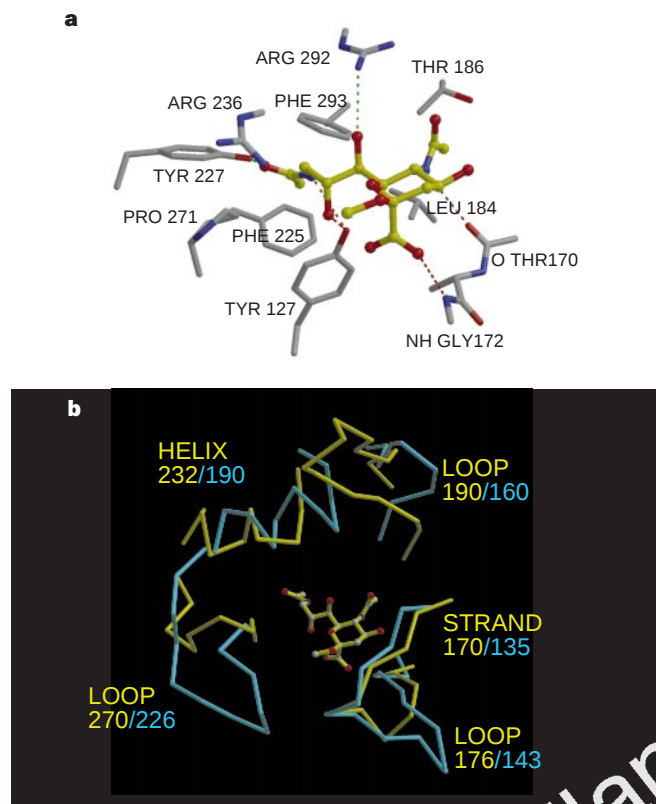


Figure 3 HEF receptor binding. **a**, Ligand bound to the receptor binding site. Potential hydrogen bonds are indicated in green or red for those conserved in HA ligand binding. Four polar contacts are formed with the ligand identically in HEF and HA: two from the hydroxyl group of HEF1 Tyr127 (Y98 in HA1) to the 8-hydroxyl and 9-amide of the ligands and two from main-chain atoms: the carbonyl oxygen of HEF1 residue 170 (135 in HA1) to the 5-amide of the ligand and the amide of HEF1 172 (137 in HA1) to the carboxylate of the ligand. The acetyl methyl group binds in a nonpolar pocket unique to HEF, formed by Phe225 and 293, and Pro271; the acetyl carbonyl oxygen contacts the hydroxyl group of Tyr224 and the guanidino group of Arg236. **b**, Comparison of architecture of the HEF and HA binding sites. The figure was constructed by superimposing the common parts of the ligands in HA and HEF.

protein into a lipophilic one, are displaced along the sequence by six residues. HEF may therefore be regarded as having an internal fusion peptide, similar to virus fusion proteins that do not require cleavage activation. The conformation of the N terminus of the HEF2 fusion peptide may indicate that fusion peptides can insert into membranes as loops.

Because the influenza C virus esterase domain E is folded like other esterases (SsEst and PAF-AH), and the R domain present in HEF and HA is a ubiquitous folding unit also found as a receptor-binding domain in the orbivirus BTV¹³, we conclude that HEF must have evolved from functional domains (Fig. 2c). A precursor to HEF may have evolved by recombination events that resulted in the insertion of R into a surface loop of E, and E into the interhelical loop of the stem domain F (Fig. 2c). In such a scheme, the trimeric F domain (Fig. 5b) may have been an ancestral membrane-fusion protein analogous to the single-function fusion proteins of paramyxoviruses such as Sendai¹⁴. Similar modular structures of the envelope glycoproteins of retroviruses are suggested by biochemical data. The first 62 and the last 20 residues of gp120 from HIV-1 can be removed, retaining receptor binding of the fragment¹⁵, indicating that those terminal segments might be analogous to F1 and F2, forming with gp41 (F3) the stem of gp160 (refs 15–17).

The HEF structure implies that the membrane fusion domains of HEF and HA consist of F1 and F2, in addition to the segment F3 = HEF2, suggesting that F1 and F2 may play a part in membrane

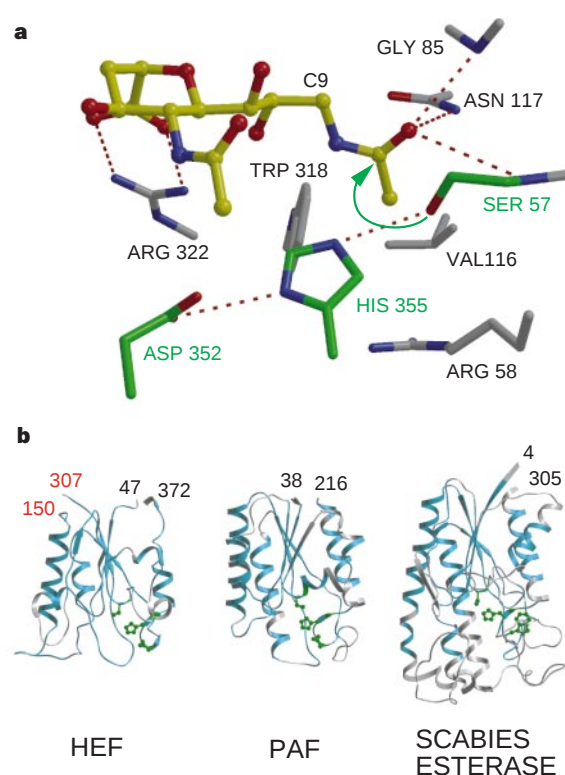


Figure 4 HEF enzyme active site. **a**, 9-Acetamidosalic acid α -methyl glycoside ($K_i = 2$ mM) or 9-acetamidosalic acid α -thiomethylmercury glycoside ($K_i = 4$ mM) shown in yellow. Catalytic triad is shown in green. Oxyanion-hole hydrogen bonds are on the right. Arg322 forms two hydrogen bonds with the sialoside carboxylate group. **b**, Esterases with structural similarity to the HEF enzyme domain (E1, E', E2; Fig. 2a). PAF (platelet-activating factor acetylhydrolase Ib) (Z-score = 8.0, sequence identity 13%, 129 aligned residues) and SsEst (esterase from *Streptomyces scabies*) (Z-score = 7.9, sequence identity 10%, 146 aligned residues). Side chains for the catalytic triads are shown in green. N and C termini of the domains are numbered (black). The HEF receptor-binding domain (R) inserts at residues marked in red. Blue ribbon indicates similar structure identified by the Dali program²⁴. HEF and PAF-AH lack the extended Ω -loops of SsEst (bottom).

fusion, either by controlling the low-pH-induced conformational change required for fusion or during the formation of a fusion pore. In both HEF and HA, F2 packs against the interhelical loop of F3, which refolds to a helix at low pH, but the position of F2 after refolding is unknown. The β -hairpin of F1 (residues 15–32 of HEF1; Fig. 2a) has already been implicated in the low-pH-induced conformational change of HA by proteolytic susceptibility¹⁸, and by its location adjacent to the site of the helix-to- β -turn refolding and chain-direction reversal on the long HA2 helix¹².

Methods

Structure determination. HEF was obtained by bromelain digestion of C/Johannesburg/1/66 virions and two crystal forms were characterized as described previously¹⁹. Crystals in harvest buffer (60% saturated ammonium sulphate, 50 mM Tris-HCl pH 7.1, 140 mM NaCl) were transferred to a ligand soak solution (40 mM 9-acetamidosalic acid α -thiomethylmercury glycoside, 300 mM MOPS pH 7.1, 140 mM NaCl) for 2 h and then cryoprotected by serial transfer through ligand soak solution containing 5–25% glycerol in 5% steps. Complexes of 9-acetamidosalic acid α -methyl glycoside ($K_i = 2$ mM) with HEF were prepared by soaking form I crystals in 40 mM ligand using the same procedure but with a slightly different soak solution (40 mM 9-acetamidosalic acid α -methyl glycoside, 100 mM Tris-HCl pH 7.1, 140 mM NaCl). 9-acetamidosalic acid α -methyl glycoside was a gift from J. Hanson.

Data collection and processing. X-ray diffraction data were collected at the Cornell High Energy Synchrotron Source by flash-cooling crystals and

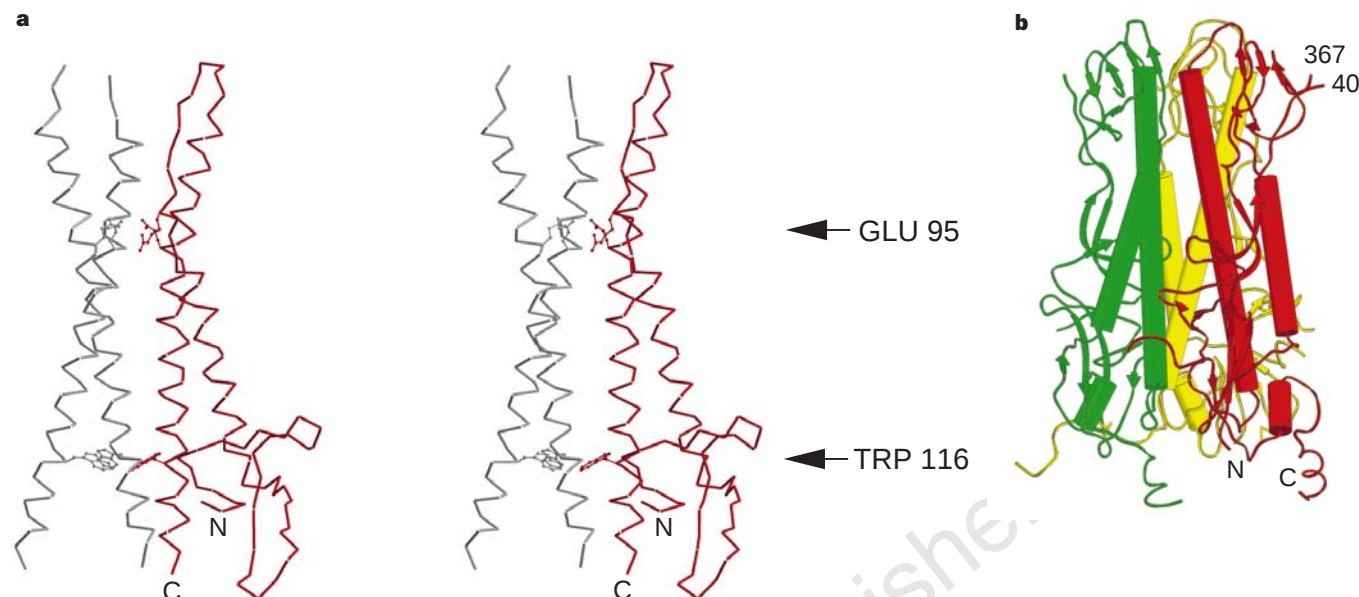


Figure 5 α -Helical interactions and the fusion peptide in HEF2. **a**, Stereo diagram of the triple-stranded α -helical bundle of HEF (79–126 of HEF2). Only the region 98–113 interacts across the trimer axis. The red monomer shows residues 4–126

(labelled N, C). **b**, Trimeric fusion domain of HEF consisting of segments F1, F2, and F3. N terminus of F1 and C terminus of F3 are indicated.

phosphorimage plate detection as described¹⁹. A 6.5 Å derivative dataset of form I crystals complexed with 9-acetamidodisialic acid α -thiomethylmercury glycoside collected on a Mar scanner and Elliot GX-13 rotating anode provided initial phases. Data were processed using Denzo, Scalepack²⁰, and programs from the CCP4 suite²¹.

Phasing, model building and refinement. 6.5 Å resolution SIR phases were extended to 3.5 Å resolution in form I crystals by iterative solvent flattening, histogram matching, and non-crystallographic symmetry averaging about the molecular three-fold axis using the program DM²². Details of the phase extension, model building, two-crystal-averaging, and refinement to 3.2 Å resolution will be described elsewhere (X.Z. *et al.*, manuscript submitted). The current model ($R_{\text{free}} = 26.7\%$, $R_{\text{work}} = 22.3\%$) contains HEF1 residues 1–427 (out of 432), residues 4–165 (out of 175) of HEF2, and no solvent molecules.

Core oligosaccharide (MAN-NAG-NAG) was built at 5 of 8 potential N-linked glycosylation sites (indicated in Fig. 2). No evidence for CHO was found at HEF1 Asn 117, occurring at the rarely glycosylated sequence NWSP. For the liganded HEF structures, the HEF model was subjected to rigid body and positional refinement against ligand data to 3.5 Å. Ligands were built into iteratively averaged $2F_o - F_c$ density phased with the HEF model, omitting residues within 5 Å of the binding sites.

Structure analysis. A Go plot²³ was used to help identify domains in HEF. The Dali program and database were used to find structurally similar domains²⁴. Least-squares superpositions were performed using the program Lsqman²⁵. RIBBONS²⁶ was used to produce Fig. 1a; GRASP²⁷ for Fig. 1b; BOBSRIPT²⁸ and Raster3D²⁹ for Figs 3a and 4a; and SETOR³⁰ for Figs 2a, 3b, 4b and 5.

Received 2 July; accepted 4 September 1998.

- Herrler, G. & Klenk, H.-D. Structure and function of the HEF glycoprotein of influenza C virus. *Adv. Virus Res.* **40**, 213–234 (1991).
- Wei, Y. *et al.* A novel variant of the catalytic triad in the *Streptomyces scabies* esterase. *Nature Struct. Biol.* **2**, 218–223 (1995).
- Ho, Y. S. *et al.* Brain acetylcholinesterase that inactivates platelet-activating factor is a G-protein-like trimer. *Nature* **385**, 89–93 (1997).
- Luytjes, W., Bredendijk, P. J., Noten, A. F., Horzinek, M. C. & Spaan, W. J. Sequence of mouse hepatitis virus A59 mRNA 2: indications for RNA recombination between coronaviruses and influenza C virus. *Virology* **166**, 415–422 (1988).
- Rosenthal, P. B. *Crystallographic Studies of the Influenza C Virus Surface Glycoprotein*. Thesis, Harvard Univ., 1996.
- Herrler, C. *et al.* A synthetic sialic acid analogue is recognized by influenza C virus as a receptor determinant but is resistant to the receptor-destroying enzyme. *J. Biol. Chem.* **267**, 12501–12505 (1992).
- Fitz, W., Rosenthal, P. B. & Wong, C.-H. Synthesis and inhibitory properties of a thiomethylmercuric sialic acid with application to the X-ray structure determination of 9-O-acetylsialic acid esterase from influenza C virus. *Bioorg. Med. Chem.* **4**, 1349–1353 (1996).
- Eisen, M. B., Sabesan, S., Skehel, J. J. & Wiley, D. C. Binding of the influenza A virus to cell-surface

receptors: Structures of five hemagglutinin-sialyloligosaccharide complexes determined by X-ray crystallography. *Virology* **232**, 19–31 (1997).

- Weis, W. *et al.* Structure of the influenza haemagglutinin complexed with its receptor, sialic acid. *Nature* **333**, 426–431 (1988).
- Rogers, G. N. *et al.* Single amino-acid substitutions in influenza haemagglutinin change receptor binding specificity. *Nature* **304**, 76–78 (1983).
- Rogers, G. N., Herrler, G., Paulson, J. C. & Klenk, H.-D. Influenza C virus uses 9-O-acetyl-N-acetylneuraminic acid as a high affinity receptor determinant for attachment to cells. *J. Biol. Chem.* **261**, 5947–5951 (1986).
- Bullough, P. A., Hughson, F. M., Skehel, J. J. & Wiley, D. C. Structure of influenza haemagglutinin at the pH of membrane fusion. *Nature* **371**, 37–43 (1994).
- Grimes, J., Basak, A. K., Roy, P. & Stuart, D. The crystal structure of bluetongue virus VP7. *Nature* **373**, 167–170 (1995).
- Skehel, A. & Choppin, P. Identification of biological activities of paramyxovirus glycoproteins. Activation of cell fusion, hemolysis and infectivity by proteolytic cleavage of an inactive precursor protein of Sendai virus. *Virology* **57**, 475–490 (1974).
- Pollard, S. R., Rosa, M. D., Rosa, J. J. & Wiley, D. C. Truncated variants of gp120 bind CD4 with high affinity and suggest a minimum CD4 binding region. *EMBO J.* **11**, 585–591 (1992).
- Wyatt, R. *et al.* Analysis of the interaction of the human immunodeficiency virus type 1 gp120 envelope glycoprotein with the gp41 transmembrane glycoprotein. *J. Virol.* **71**, 9722–9731 (1997).
- Helseth, E., Olshefsky, U., Furman, C. & Sodroski, J. Human immunodeficiency virus type 1 gp120 envelope glycoprotein regions important for association with the gp41 transmembrane glycoprotein. *J. Virol.* **65**, 2119–2123 (1991).
- Skehel, J. J. Changes in the conformation of influenza virus hemagglutinin at the pH optimum of virus-mediated membrane fusion. *Proc. Natl. Acad. Sci. USA* **79**, 968–972 (1982).
- Rosenthal, P. B. *et al.* Crystallization and preliminary X-ray diffraction studies of the influenza C virus glycoprotein. *Acta Crystallogr. D* **52**, 1041–1045 (1996).
- Otwonowski, Z. & Minor W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
- Collaborative Computational Project No. 4, the CCP4 Suite: Programs for Protein Crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Cowan, K. D. *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography* **31**, 34–38 (1994).
- Go, M. Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature* **291**, 90–92 (1981).
- Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138 (1993).
- Kleywegt, G. J. & Jones, T. A. *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography* **31**, 9–14 (1994).
- Carson, M. Ribbon models of macromolecules. *J. Mol. Graph. Model* **5**, 103–106 (1987).
- Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct. Funct. Genet.* **11**, 281–296 (1991).
- Esnouf, R. M. An extensively modified version of Molscript which includes greatly enhanced coloring capabilities. *J. Mol. Graph. Model* **15**, 132–134 (1997).
- Merritt, E. A. & Bacon, D. J. Raster3D: photorealistic molecular graphics. *Methods Enzymol.* **276**, 505–524 (1997).
- Evans, S. V. SETOR: Hardware-lighted three-dimensional solid model representations of macromolecules. *J. Mol. Graph. Model* **11**, 134–138 (1993).

Acknowledgements. We thank M. Frayser and R. Crouse for technical assistance, and members of the Harrison–Wiley laboratory and the staff of the Cornell High Energy Synchrotron Source for assistance with data collection. P.B.R. received graduate research assistant support from the Department of Molecular and Cellular Biology and the committee on higher degrees in biophysics, Harvard University, as a student, and an NIH training grant as a postdoctoral fellow; X.Z. was supported by the HHMI. This work was supported by the NIH, the Deutsche Forschungsgemeinschaft, the MRC and the HHMI. D.C.W. is an investigator of the HHMI.

Correspondence and requests for materials should be addressed to D.C.W. Coordinates will be deposited in the Brookhaven Database and are available before release from <http://www.crystal.harvard.edu>.